Developing Methods for Causal Mediation Analysis of Parenting Interventions to Improve Child Antisocial Behaviour

Zhang, Cheng

Awarding institution:
King's College London

Developing Methods for Causal Mediation Analysis of Parenting Interventions to Improve Child Antisocial Behaviour

Cheng Zhang

A thesis submitted to the King's College London for the degree of Doctor of Philosophy

Supervisors:

Professor Sabine Landau

Professor Stephen Scott

Department of Biostatistics

Institute of Psychiatry, Psychology & Neuroscience

King's College London

March 2015

Abstract

Parenting programmes are the most effective intervention to change persistent child antisocial behaviour and are widely used, but little is known about the mechanisms through which they work and hence how to improve them. This PhD project aims to bridge this gap by performing formal mediation analyses partitioning total effects of parenting programmes on child outcome into indirect effects (mediated through aspects of parenting) and direct effects (non-mediated effects). This thesis focuses on further developing methods for mediation analysis to cover complex scenarios and applies them in three trials (SPOKES, CPT and HCA) of parenting programmes.

This project improves traditional methods for trials that assume no putative mediator-outcome confounding in three ways: Firstly, the mediator-outcome relationship is adjusted for observed confounding variables. The newly developed MI-BT method facilitates the application of Multiple Imputation to handle missing data and the use of linear mixed models to reflect trial design, and generates non-parametric inferences via a bootstrap approach. The application of this method to the SPOKES trial showed statistically significant indirect effects for two mediators (parental warmth and criticism). Secondly, the MI-BT method is extended to combine with instrumental variables method and become the IV-MI-BT method which allows for unmeasured confounding of the mediator-outcome relationship in the presence of missing data. The application of this method to the SPOKES trial showed that while IV estimators of mediation effects were similar in value compared to MI-BT estimates, their confidence intervals were inflated. Finally, methods were further developed to enable pooling of individual participant data from multiple trials and so provide for potentially more precise and more generalizable mediation analyses. A framework for systematically conducting such an IV-MI-BT IPD meta-mediation analysis is described. Meta-analysis of the three contributing trials did not detect any evidence for between-trial heterogeneity in mediation effects of interest. Pooling of the studies resulted in smaller and non-significant overall indirect effect estimates and provided a considerable precision gain compared to the SPOKES only analysis.

Table of Contents

List of Tables

List of Figures

Acknowledgement

I would like to thank my supervisors, Professor Sabine Landau and Professor Stephen Scott, for the patient guidance, encouragement and advice they have provided throughout my time as their student. Without their persistent help this thesis would not have been possible. It has been a pleasant experience working with my supervisors who cared so much about my work, and who responded to my questions and queries so promptly.

I would also like to thank Biostatistics Department of King's and Biomedical Research Centre for giving me the opportunity to write this PhD thesis. To Professor Andrew Pickles, I am very grateful for his assistance and suggestions throughout my project. To all staffs at Biostatistics Department of IOPPN, thank you very much for always encouraging me and helping me survive all the stress. I have been extremely lucky to have colleagues who have been so caring.

Finally, I am indebted to my family and friends for their support. Special thanks to my husband who is always listening and understanding. It has been a memorable three and a half years to both of us undergoing an adventure of our PhDs.

Notation

Observed variables:

R: Randomly allocated treatment group (R=1 for active treatment; R=0 for control)

Y: Observed distal outcome (measure of child antisocial behaviour in this project)

M: Observed intermediate outcome (aspect of parenting in this project)

W: Observed putative moderator of the effect of treatment on outcome

$\mathbf{X}$: Observed baseline confounders

$\mathbf{Z}$: Observed instrumental variables

Potential (counterfactual) outcomes:

$M(1)$: Intermediate outcome if participants were randomised to active treatment group

$M(0)$: Intermediate outcome if participants were randomised to control condition

$Y(r, m)$: Outcome with intervention r and mediator m

$Y(1) = Y(1, M(1))$: Outcome if participants were randomised to active treatment group with intermediate outcome $M(1)$

$Y(0) = Y(0, M(0))$: Outcome if participants were randomised to control condition group with intermediate outcome $M(0)$

Acronyms:

ETM: Effect of Treatment on Mediator

EMO: Effect of Mediator on Outcome

IE: Indirect Effect

DE: Direct Effect

TE: Total Effect

Chapter 1 General Introduction

Causal mediation analysis plays a key role in the investigation of mechanisms in the social and behavioural sciences. In particular, mediation analysis for psychological intervention trials is of great current interest to researchers studying the mechanisms through which complex interventions improve outcomes. In child mental health, parenting programmes have been developed based on the theory that these types of complex interventions improve parents' parenting practices, which in turn should reduce children's antisocial behaviours. Many Randomised Controlled Trials (RCTs) of parenting programmes have been conducted in recent decades, and several study-level meta-analyses have been used to synthesise evidence from comparative effectiveness studies (NICE, 2013). Although the effectiveness of parenting programmes is well established, formal mediation analyses for investigating the mechanisms by which improvements in child outcomes are brought about are still in their infancy. This PhD project aims to bridge this gap and develop methods for mediation analysis in trials of parenting interventions. I then apply these methods to investigate mediation using data from three existing trials of the "Incredible Years" (IY) parenting programme.

The first chapter of this thesis begins by reviewing the theory underlying the development of parenting interventions (Section 1.1). It will then go on to describe the existing parenting programmes for improving child antisocial behaviour and the evidence for their effectiveness (Section 1.2). After laying out previous mediation investigations in studies of parenting interventions (Section 1.3), I will provide some background regarding the statistical challenges posed by mediation analysis (Section 1.4). The outline of the thesis will be stated at the end of this chapter (Section 1.5).

1.1 Child antisocial behaviour

Persistent antisocial behaviour in children is common, costly and often leads to poor outcomes in adulthood. The UK Office for National Statistics reported that 5.8% of children aged 5 to 15 years had clinically significant conduct problems using the relatively conservative ICD-10 (World Health Organization, 1992) research criteria for conduct disorders (Meltzer et al., 2000). Children with this pattern of behaviour are often rejected by their peers, are more likely to become involved in a deviant peer group, most leave school

with no qualifications, and they are at high risk for engaging in chronic delinquent behaviours. Antisocial behaviour usually continues from childhood to adulthood, with long-lasting difficulties in relationships with other people, and the employment pattern is poor (Fergusson et al., 2005). Individuals with conduct disorder aged 10 cost up to ten times more than controls by age 28, and these costs involve many agencies (Scott et al., 2001a).

The well-established social learning theory and the attachment theory (Scott and Dadds, 2009) both provided the essential theoretical foundation for the development of parenting training interventions aimed at improving the parent-child relationship. In combination with supportive empirical findings, these two systematic parent-child theories have become the guide to establishing hypothesised models that researchers can apply to investigate the mechanism of parenting interventions. The following sections provide a general overview of social learning theory, attachment theory and child behaviour outcome research findings for understanding the underline theoretical mechanism of parenting interventions and offering guidance as to the possible mediating factors in parenting interventions to improve child antisocial behaviour.

1.1.1 The contribution of social learning theory

A wealth of empirical studies based on social learning principles provides strong evidence linking child antisocial behaviour to poor quality parenting. Child antisocial behaviours (also known as child conduct problems) typically include: troublesome, disruptive and aggressive behaviour; an unwillingness/inability to perform school work; few positive interactions with adults; poor social skills; non-compliance with instructions, and emotional volatility (Youngstrom et al., 2000). The parenting styles associated with these outcomes include high criticism and hostile parenting, harsh punishment, inconsistent discipline, low warmth, low involvement, low encouragement, and poor supervision (Denham et al., 2000).

Patterson and colleagues' clinical investigations support the view that parents inadvertently directly train the child to perform antisocial behaviours (Patterson, 1982, Patterson et al., 1989). The parents tend to be non-contingent in their use of positive reinforcers. Simply put, the child's prosocial actions are ignored or inappropriately responded to, and the child's antisocial behaviours are non-contingently rewarded or ineffectively punished, leading to mutually coercive cycles of interaction between parents and child. The parents directly

reinforce the child's antisocial behaviours by applying both positive reinforcement and negative reinforcement without being aware of doing so. Positive reinforcement such as giving attention (even with a disapproving comment) can strengthen the probability that the child's antisocial behaviour will be repeated in future, since the child has learned that he or she is gaining attention.

Probably the most important set of contingencies for antisocial behaviour are escape-conditioning contingencies, in which the child uses aversive behaviours to terminate aversive intrusions by the parents. For example, during a conflict bout, the parents give in after the child's persistent or escalating antisocial behaviours and the child wins. As the moment-to-moment exchange goes on, the child and parents gradually escalate the intensity of their coercive behaviours, sometimes including hitting and physical attacks. Eventually, the child learns to control the parents through coercive means. In parallel, the child fails to learn social skills. If parent-child interactions and children's exposure conditions do not change, the learning process will carry on, and when children grow older, they will tend to apply what they have learnt to an extended range of people and situations.

In numerous longitudinal studies, the parenting practices of parents of antisocial children have been characterized by harsh and inconsistent discipline, little positive parental involvement with the child, and poor monitoring and supervision of the child's activities (Loeber and Dishion, 1983). Although the magnitude of predictive power varies from study to study, these sorts of parenting practices are clearly identified as a major risk factor for children's antisocial behaviour. In the UK Cambridge study, a large-scale prospective study, Farrington and colleagues (Farrington, 1995) found that parental discipline was identified as one of the best independent predictors at early childhood (age 8-10) of adolescent aggression and also adult violence (up to age 32).

Based on social learning theory, interventions targeted at changing parenting practices were established for reducing child antisocial behaviours. In carefully designed trials of this type of intervention, parents were randomly assigned to intervention (e.g. parent training) and control groups. Results of trials of parenting programmes and several meta-analyses (Martinez and Forgatch, 2001, Scott and Dadds, 2009) showed an improvement of parenting practices and a reduction of child antisocial behaviour. This finding suggests the existence of

a linkage between parenting practice and child antisocial behaviour. Trials of parenting intervention provide rich information to enable researchers to investigate the causal relationship between intervention, parenting practices and child outcome using appropriate statistical methods. More details of parenting interventions for improving child antisocial behaviour are described in Section 1.2 and a review of statistical mediation analyses on parenting trials is included in Section 1.3.

In summary, parenting practices as formulated by social learning theory are highly correlated with child antisocial behaviour, and parenting interventions focused on improving parenting practices can lead to a decrease in child antisocial behaviour. As child antisocial behaviour is not specifically associated with a single dimension of the parent-child relationship, parenting programs do not usually target only one aspect of the parent-child relationship. Moving beyond parental disciplinary practices alone, effective parenting practices need to be embedded in a positive, responsive and supportive parent-child relationship (Dishion et al., 1996). While discipline practices and parental behaviours and beliefs based on social learning theory have been critical to developing effective interventions, the wider content of the parent-child relationship, including its emotional aspects, has been helpfully illuminated by attachment theory.

1.1.2 The contribution of attachment theory

As one of the most influential models of parent-child relationships, attachment theory draws upon a broad range of theoretical fields, building on concepts from ethology, cognitive psychology and control systems (Bowlby, 1969/1982, Ainsworth et al., 1978). The theory has now been broadly applied in the fields of social and emotional development, illuminating clinical problems on early parent-child relationships, and has led to supporting research on close relationships of adolescents and adults (Cassidy and Shaver, 2008). John Bowlby is the father of attachment theory, in which he explained the nature, significance and function of the bonds of affection between children and their caregivers and the impact of their disruption through separation, deprivation and bereavement. Attachment theory proposed that infant behaviour associated with attachment is primarily the seeking of proximity to an attachment figure, especially under organismic (e.g. hunger, fatigue, illness and unhappiness) and environmental (e.g. strange situation) distress, through which the infant gains more protection and increases his or her probability of survival. The infant forms

attachments to a consistent caregiver who is sensitive and responsive to the infant's signalling behaviours and who engages the infant in social interaction. The attachment figure is used as a secure base to explore the world and as a safe haven to return to in times of trouble. It is postulated that attachment relationships play a crucial role in the child's development of attachment patterns and internal working models, which will carry forward to influence the child's capacity for building relationships with others (Bretherton and Munholland, 1999). Continued disruption of the attachment between infant and primary caregiver can result in long term behaviour, cognitive, social, and emotional difficulties for that infant.

Mary Ainsworth and colleagues in the 1960s and 1970s developed the Strange Situation Procedure, which is a widely used method of assessing an infant's pattern and style of attachment to a primary caregiver. In the Strange Situation Procedure, young children's reactions relative to their caregivers were observed in a series of separations and reunions, and the children were categorized into four groups (Avoidant, Secure, Resistant and Disorganized) based on their reactions. Each of these groups reflects a different kind of attachment style with the caregiver. Secure (B) babies are securely attached to their attachment figure, using the caregiver as a secure base for exploration and seeking help. Infants develop a secure attachment when the caregiver is sensitive to their signals, and responds appropriately to their needs. Avoidant (A) babies avoid or ignore the caregiver. The babies' needs are frequently not met and they think themselves unworthy and unacceptable, due to a rejecting primary caregiver (Larose and Bernier, 2001). Avoidant babies are more likely to be hostile, aggressive, disruptive pre-schoolers (Greenberg et al., 1993). Resistant (C) babies adopt an ambivalent behavioural style towards the caregiver and this attachment style is associated with inconsistent primary care and indicates a greater likelihood for attachment (internalizing) problems in the future (Kobak et al., 1993). Disorganized (D) babies' behaviours are disoriented and unpredictable. This insecure attachment style is strongly related to risk for childhood psychopathology, especially antisocial behaviours (Greenberg, 1999).

Attachment theory emphasizes that a secure attachment relationship between a mother and child is a protective factor against the development of child antisocial behaviour. Conversely, insecure attachment styles (Avoidant, Resistant and Disorganized) have been found to

adversely affect behaviour in childhood. The theoretical underpinnings of this relationship highlight the importance of parental emotion (i.e. expressed warmth and criticism). Greenberg's study suggested that dismissive parenting with a lack of warmth is more likely to lead to insecure patterns that are associated with antisocial behaviour (Greenberg, 1999). Children whose parents have low warmth and a hostile parenting style show significantly worse outcomes than children whose parents have warmth and a positive control parenting style (Baumrind, 1991, Hetherington, 1999). Additionally, it is worth pointing out that "parent-child relationship" defines a broader scope than "parent-child attachment", as the former involves not only emotional security but also discipline, cognitive stimulation and condition etc. (Scott, 2008). Therefore, effective parenting practice interventions for reducing child antisocial behaviour have been established to improve both parental disciplinary practices and emotional aspects such as parental warmth and criticism.

1.2 Parenting programmes for improving child antisocial behaviours

Early intervention is important to prevent high risk children developing poor outcomes. It is also potentially a cost-effective method in the battle against child antisocial behavioural problems. Numerous interventions have been developed, including behaviour therapy, residential treatment, family therapy and parental skill interventions that target early-starting, high-risk children. The most theoretically and empirically well-established treatment for child antisocial behaviours is Parent Management Training and parent training programmes have grown rapidly in a number of countries over recent decades (Kazdin, 2005).

1.2.1 Typical parent training interventions

Parent training interventions have been developed over forty years and the more cost-effective ones are characterised as structured, short term (mostly 6 – 20 weeks) and delivered in a variety of settings (home, school, clinic or community) and media (telephone, video or face-to-face). They are directed at parents to improve parenting skills and enhance the parent-child relationship, and should consequently reduce child problem behaviours. Well known programmes include the "Incredible Years" Parenting Programme (Webster-Stratton and Hancock, 1998), the Triple P Parenting Programme (Sanders, 1999), Parent–

Child Interaction Therapy (Bagner and Eyberg, 2007) and the Parent Management Training Oregon model (Forgatch et al., 2005).

The "Incredible Years" (IY) Parenting Programme is one of the most internationally recognizable parent training programmes. It was developed by Webster-Stratton (Webster-Stratton and Hancock, 1998). The programme attempts to provide parent training to strengthen the parent's competencies in monitoring and appropriately disciplining the child's behaviours and to increase the parent's involvement in the child's experiences (i.e. enhance positive parenting and reduce negative parenting), and therefore to promote the child's social skills and reduce the child's antisocial behaviours. Typically in this intervention, trained experts use parent training videotapes to teach groups of parents more appropriate ways of handling children. There are a series of tailored and age-appropriate versions of the programme.

The Triple P Positive Parenting Programme is a comprehensive, multi-level prevention programme originally developed by Sanders (Sanders, 1999). The main purpose of the programme is to prevent severe behavioural, emotional, and developmental problems in children by enhancing the knowledge, positive parenting skills, and confidence of parents. The program is administered at five different levels, depending on the severity of the child's behavioural problems. Level 1 provides a universal parenting information strategy. Level 2 offers guidance and advice to parents of children with mild behaviour problems. Level 3 is a four-session parent active skills training program that targets children with mild to moderate behaviour problems, and Level 4 is an intensive parent training programme for children with serious behaviour problems composed of eight to ten sessions. Finally, Level 5 is an enhanced programme for families where parenting difficulties are combined with other sources of family distress.

The Parent–Child Interaction Therapy (PCIT) is an individually-delivered and not group-based parent training program. It aims to promote the child's appropriate behaviours and reduce the child's antisocial behaviours by fostering a responsive parent-child relationship (Bagner and Eyberg, 2007). The intervention program is usually organized in two phases: The child-directed interaction phase (CDI) attempts to enhance the parent–child relationship, increase positive parenting, and improve child social skill, while the parent-directed interaction phase (PDI) focuses on training the parents how to set limits, reward properly the child's

compliance and punish noncompliance. In the PCIT programme, the trained therapists use instruction, modelling, and various role-playing techniques to coach parents toward mastery of the interaction skills.

The Parent Management Training Oregon model (PMTO) is a Social Interaction Learning (SIL) theory based approach developed by the group at the Oregon Social Learning Centre (OSLC: (Forgatch et al., 2005). It is a short-term family intervention delivered directly to the parents by the trained therapists according to PMTO manuals. Families are treated individually, so that the intervention is flexible enough to meet the individual family's needs. The aim of the PMTO is to enhance five core effective parenting skills: Encouragement, Discipline, Monitoring, Problem-solving and Positive involvement. PMTO has been tailored for clinical problems and prevention designs associated with children's antisocial behaviour. Several manuals detail related procedures for various contexts or focal populations.

1.2.2 Measurement of child behaviours

Multi-informant measurement of child behavioural outcomes is common in parenting training programme trials. Parent report of outcome, teacher report of outcome and direct observation by independent observers are used for assessing child behaviours. Multiple methods such as questionnaires, interviews and direct observation are adopted for the collection of child outcome information. A number of standardised instruments have been developed for this need and some examples are listed below.

Questionnaires - The Strengths and Difficulties Questionnaire (SDQ) is a one-page questionnaire for assessing the pro-social behaviour and psychopathology of 3–16-year-olds. It can be completed by parents, teachers, or youths (Goodman, 2001). The SDQ consists of twenty-five positive or negative attributes rated on a three-point Likert scale to indicate how far each attribute applies to the target child. The scores for emotional symptoms, conduct problems, hyperactivity-inattention, peer problems, and pro-social behaviour are generated by dividing the twenty-five items into five scales and the total difficulties score is the sum of all the scores. The SDQ can be used for screening, as part of a clinical assessment, as a treatment-outcome measure, and as a research tool. Similarly, the Child Behaviour Checklist (CBCL) is a device by which parents or other individuals who know the child well rate a child's problem behaviours and competencies. The questionnaire consists of 20 competence items

and 120 items on behaviour or emotional problems (Achenbach and Rescoral, 2001). The Eyberg Child Behaviour Inventory (ECBI) is a parental report of conduct behavioural problems in children and adolescents that measures the intensity (the number of difficult behaviour problems) and the frequency with which they occur (Eyberg and Pincus, 1999).

Interviews - The Parental Account of Children's Symptoms (PACS) is a standardised, semi-structured interview that was developed as an instrument for the measurement of children's behaviour problems as seen at home (Taylor et al., 1986). Parents are asked by a trained interviewer for detailed descriptions of what their child has done in specified situations over the previous week. Ratings (frequency and severity) are made by the interviewers based on their training and written definitions of the behaviours. The final score is constructed by taking the average frequency score and severity score with a continuous range from 0 to 3. One well-known semi-structured interview is the Child and Adolescent Psychiatric Assessment (CAPA), which is designed to assess psychiatric symptoms occurring during the preceding three-month period in youths aged between 9 and 17 (Angold and Costello, 2000). The CAPA is a 1-2 hour interview that collects data on the onset dates, duration, frequency, and intensity of psychiatric symptoms and also includes an assessment of psychosocial impairment and clinician ratings of behaviours observed in the interview.

Independent Direct Observations - The Dyadic Parent Interactive Child Scale (DPICS) is designed for use in assessing the quality of parent-child social interaction. The parent and child are observed in three standard situations (Child-Directed Interaction, Parent-Directed Interaction and Clean-up) that vary in the degree to which parental control is required. There are twenty-four standard parent and child behaviour categories which are coded during observation of each situation (Robinson and Eyberg, 1981). A similar system has been developed by the CPPRG (Conduct Problems Prevention Research Group, 1999). Observational assessments of parenting and child behaviour are performed in the three tasks, including “free play” (child-led, 5 minutes), “a parent control situation” (parent-led, 5 minutes), and “tidy-up” (5 minutes). However, direct observation is not recommended to assess antisocial behaviour, since the paradigm used (one-to-one activity with the parent) fails to elicit substantive oppositional behaviour and is not reliably predictive of current or later disruptiveness (Wakschlag et al., 2008).

1.2.3 Measurement of parenting practices

Similar to the measurement of child behaviours in parenting programme trials, parenting outcomes are assessed using multiple informants and multiple methods. Questionnaires, interviews and direct observation are employed for measuring parenting practices.

One of the standardised questionnaires is the Parent Practices Questionnaire (PPQ). It has four subscales: positive involvement, appropriate discipline, inconsistent parenting and harsh discipline (Webster-Stratton et al., 2008). Additionally, the Alabama Parenting Questionnaire (APQ: (Shelton et al., 1996) is a fifteen-item scale measuring parenting behaviour, consisting of five subscales made up of three items each: “Positive Parenting”, “Inconsistent Discipline”, “Poor Supervision”, “Involvement” and “Corporal Punishment”. Each item is measured across a scale from 1 to 5 from “Never” to “Always”, giving a total possible score ranging between 3 and 15 for each of the five subscales.

Michael and colleagues (Dowdney et al., 1985, Dowdney et al., 1984) developed a semi-structured interview that measures the frequency of the withdrawal of the child’s privileges, the frequency with which the child is praised or rewarded and the frequency of implementation of “timeout” punishment or harsh discipline (smacking). These single items are scored on a scale of 0 to 4. As a subscale relating to the frequency of time the parent spends playing with the child, creative play is derived by taking the mean score of three play tasks (pretend play, Lego and drawing) and scored from 0 to 4 on a continuous scale.

Expressed emotion (EE): this is a measure of emotions expressed towards the child throughout the interview using Camberwell Family Interview criteria (Vaughn, 1989, Brown et al., 1972). Expressed warmth is rated on a four-point scale: 0=no expressed warmth, 1=some warmth, 2=moderate warmth and 3=a great deal of expressed warmth. Expressed criticism is rated on a five-point scale: 0=no criticism, 1=very little criticism, 2=moderate criticism, 3=quite a lot of criticism and 4=a lot of criticism throughout.

As mentioned in section 1.2.2, the CPPRG developed a direct observation system to videotape the parent-child interaction for fifteen minutes across three tasks and the frequency counts of parent behaviours of each category are rated using CPPRG scoring scheme. The average score is calculated by taking the mean over three tasks for each scale.

For example, the scale of “average parent’s negative affect” is derived by averaging the values measured among the “free play”, “parent led play” and “tidy up” tasks. A subscale is also derived for measuring positive parenting using the averaged scales, i.e. the “parent’s total attention” is calculated as the sum of “average attentiveness”, “average positive attention”, and “average praise”.

1.2.4 Strengths and weaknesses of different measurement methods

Different measurement methods have their strengths and weaknesses in measuring parent and child outcomes in practice. Questionnaires have reasonable psychometric properties, but are subject to social desirability bias: that is, the tendency of respondents to answer questions in a manner that will be viewed favourably by others. As questionnaires are completed by parents themselves, the same outcome may be rated differently to a large extent by different parents due to the differences in their attitude towards children. To reduce this desirability bias, parents are asked to provide detailed examples and this information is rated by trained investigators. In fact, this is the key feature of interviews: they use objective investigator-based criteria based on detailed descriptions of the problems rather than parental impressions. In terms of direct observation measurement methods, they are totally objective, but the disadvantage is that they are very short periods of observation that may not be typical of what is going on at home. Table 1-1 compares three parenting programme outcome measurement methods - questionnaire, interview and direct observation - in terms of cost, training duration, time coverage, opinion source and social desirability bias. Employing the multi-method approach combines the strengths of each method and reduces the impact of the weaknesses of an individual measurement method.

Table 1-1 Parenting Programme Outcome Measurement Methods Comparison

	Measurement Methods		
Domain	Questionnaire	Interview	Observation
Cost	Low	Moderate	High
Investigator Training Duration	None	1 month	2-3 months
Opinion Source	Parent	Parent information with investigator rating	Objective -Independent rating
Time Coverage	Preceding 6 months or a year	Previous 6 months or a year	Usually 1-2 hours
Risk of Social Desirability Bias	High	Moderate	Low

1.2.5 Evidence of effectiveness of parenting programmes

Systematic review results of randomised controlled experimental studies indicate that early family/parent training is an effective intervention for reducing behaviour problems among young children. Based on fifty-five studies, one meta-analysis (Piquero et al., 2009) estimated the weighted standardised effect size to be 0.35 with a 95% confidence interval from 0.26 to 0.44. Meta-analysis results also indicate that larger studies (sample size >100) tend to have less effect size variations compared with studies that were based on small samples. Review results of long-term follow up studies suggest that early family/parent training is also effective in reducing delinquency and crime in later adolescence and adulthood (Farrington and Welsh, 2003).

A systematic review of RCTs of group-based parent-training programmes for the treatment of children with conduct problems concluded that group-based parenting programmes are an effective treatment for such children (Dretzke et al., 2009). Meta-analysis results of RCTs of twenty-four group-based parent-training programmes showed statistically significant differences favouring the intervention group using both parent and independent reports of child outcome. The standardised mean difference of child conduct problems between the intervention group and the control group (intervention - control) was -0.67 with a 95% confidence interval from -0.91 to -0.42 using parent reports of child outcome, and -0.44 with a 95% confidence interval from -0.66 to -0.23 using independent reports of child outcome. However, the relative effectiveness of parenting programmes requires further research because there was insufficient evidence to determine the relative effectiveness of different approaches to delivering parenting programmes.

Two reviews of the Triple P Positive Parenting Programmes focus on the effectiveness of the interventions on parenting and on behavioural problems in children respectively (de Graaf et al., 2008b, de Graaf et al., 2008a). Meta-analysis results indicated that the Triple P Level 4 interventions reduced dysfunctional parenting styles in parents, improved parental competency (de Graaf et al., 2008b) and also reduced disruptive behaviours in children (de Graaf et al., 2008a). These effects were maintained well through time and with further improvements in long-term follow-up.

A series of systematic reviews and meta-analyses included in the National Clinical Guideline Number 158 (NICE, 2013) provide the statistical effectiveness and health economic evidence of a wide range of interventions on improving antisocial behaviour in children and young people. Systematic reviews were conducted in interventions of four categories respectively, namely selective interventions, psychosocial indicated interventions and pharmacological and physical treatment interventions. For the parent-focused (delivered to parent only) psychosocial interventions, fifty-four RCTs (4,150 samples) from eleven countries were included in a meta-analysis. The meta-analysis results showed that parent-focused interventions reduced antisocial behaviour when rated by observers, researchers/clinicians and parents at post-treatment. The mean parent-rated antisocial behaviour in the intervention groups was 0.54 standard deviations lower (95% CI from 0.65 to 0.44) than in the control group. It was 0.40 standard deviations lower (95% CI from 0.58 to 0.21) for the observer-rated antisocial behaviour, and 0.69 standard deviations lower (95% CI from 1.22 to 0.16) for the researchers/clinicians-rated antisocial behaviour. However, there was no evidence of benefit when antisocial behaviour was rated by teachers.

This PhD project includes trials of three IY parenting programmes: CPT (Scott *et al.*, 2001b), SPOKES (Scott et al., 2010b), and HCA (Scott et al., 2012a). These trials provided parent-focused interventions for reducing children's antisocial behaviours that were rated by their parents. Here, I will focus on the primary findings of the trials, namely the intervention effects on reducing antisocial behaviour. Details of the design and the measurement of the trials will be reviewed in Chapter 2 of this thesis. The CPT trial found a significant reduction of child antisocial behaviour in the IY intervention group compared with the waiting list control with an effect size of 1.06 standard deviations (95% CI from 0.71 to 1.41). The results of the SPOKES trial showed that child antisocial behaviour in the intervention group is significantly lower than in the control group, with an effect estimate of -0.24 (95% CI from -0.35 to -0.12) using the PACS interview measurement. Finally, the HCA trial detected a significant interaction for the treatment groups over time (pre and post treatment) on anti-social behaviour between the active intervention groups relative to the control group. These results were controlled for the child's age at assessment and whether the child had any special needs.

In summary, I have provided evidence for the effectiveness of parenting programmes on reducing child antisocial behaviour via reviewing the results of meta-analyses of parenting intervention trials. Focused on the PhD project trials, I have showed the results of analysis of these three RCTs of the IY parenting programme on improving child antisocial behaviour outcomes. In the next section, the effects of the interventions on parenting outcomes will be discussed (Table 1-2) as a part of the review on mediation investigations in studies of parenting interventions for child antisocial behaviour.

1.3 Mediation investigations in studies of parenting programmes for child behavioural problems

Even though the effectiveness of well-established parenting programmes for child behavioural problems is widely established, the mechanisms by which interventions affect changes in children's outcomes and for whom the interventions are effective (i.e. mediation and moderation) are much less investigated. Understanding the mechanisms of the interventions is pivotal to refining their implementation, improving services delivered, and advancing theory (La Greca et al., 2009, Rutter, 2005). Investigating how interventions improve child outcomes by means of robust mediation analysis is the focus of this thesis.

1.3.1 Mediation analyses conducted in parenting programmes

Despite the rapid progress made by RCTs of parenting training programmes, mediation of intervention effects by parenting outcomes targeted by the intervention is rarely tested. Typically, the primary objective of an RCT of a parenting intervention is to test the effectiveness of the intervention in terms of improving child behaviour. Additionally, parenting skills may be measured at one or several time points as the secondary outcome (Beauchaine et al., 2005, Reid et al., 2004). In the majority of trials, interventions are formally evaluated in terms of effects on primary and secondary outcomes (which may include putative mediators); mediation is inferred if both show intervention effects but mediation is rarely formally established.

Mediation is traditionally established by demonstrating four logical relationships among treatment, mediator and outcome (Baron and Kenny, 1986, Weersing and Weisz, 2002). First, is the intervention efficacious in terms of clinical outcome? Second, is there an

intervention effect on the mediator? Third, does the mediator have a significant effect on the clinical outcome, when both the intervention effect and the mediator variable are included as predictors of the outcome variable? And fourth, is the mediation effect (the effect that goes through the causal mediation pathway) significant? (Weersing and Weisz, 2002). More recently, mediation has been assessed by simultaneously modelling the putative mediators and the outcome variable using structural equation models (SEMs; e.g. path models).

1.3.2 Summary of mediation effects assessed in RCTs of parenting training programmes for child behaviour problems

Baron and Kenny's approach to mediation analysis was used in a Swedish parent management training (PMT) RCT (Kling et al., 2010) to investigate the mediating effects of the parenting practices and homework fidelity on child conduct problems. Three regression models (intervention – mediator model; intervention – outcome model; intervention and mediator – outcome model) were fitted for each of these potential mediator variables. Two summary scores were presented for the parenting practice subscales – the summary score representing harsh and inconsistent parenting (HI); and the summary score representing praise and incentives (PI). Partial mediation was established for both the HI score and the PI score, which were significant (z=2.6, p<0.01 and z=2.0, p<0.05, respectively) according to the Sobel test (Baron and Kenny, 1986). The hypothesis that intervention effects were mediated by homework fidelity was also investigated using Baron and Kenny's approach. The analysis revealed that homework fidelity completely mediated the change in conduct problems at post-test and follow-up, which was significant according to the Sobel test (z=2.1, p<0.05). The lower the HI parenting score, the higher the PI parenting score or the more homework the parents completed, the larger the reductions of conduct problems. The statistical aspects of Baron and Kenny's approach and the Sobel test will be reviewed in section 1.4.

Similarly, based on the results of an "Incredible Years" parenting group programme RCT, Gardner *et al*. (Gardner et al., 2006) concluded that change in parenting skills appears to be a key mechanism for change in child problem behaviour. Whether observed positive parenting is a mediating mechanism for treatment change was tested by assessing correlation among three variables and fitting hierarchical multiple regression models: The results showed that improvement in observed positive parenting correlated with

improvement in observed child negative behaviour (r = 0.40, p = 0.001); treatment status correlated with change in positive parenting (r = 0.27, p = 0.03) and change in child negative behaviour (r = 0.35, p = 0.004). A significant partial mediation effect (p < 0.025) established by using the PRODCLIN programme (MacKinnon et al., 2007b) was detected, since the effect of treatment on negative behaviour was attenuated when positive parenting was introduced as a potential mediating variable. However, no significant correlation was found between child negative behaviour and parental sense of competence. In a secondary study (Gardner et al., 2010), the mediation analysis results also found that change in observed positive parenting mediated change in child negative behaviour using a regression approach in combination with the Sobel test.

Eddy and Chamberlain (Eddy and Chamberlain, 2000) used Structural Equation Modelling (SEM) to test the effects of Multidimensional Treatment Foster Care (MTFC) placement on later antisocial behaviour and to assess whether the effects of placement were mediated by behaviour management skills and deviant peer association. The analysis model assumed that family management practices and peer association were representing the same underlying latent variable. The paths from treatment to mediators, from mediators to follow-up antisocial behaviour, and from pre-antisocial behaviour to follow-up antisocial behaviour were all significant. These positive effects of MFTC on behaviour were mediated by the latent variable capturing the combined effect of behaviour management and peer association (z= -2.72, p <0.05) (Sobel, 1987) and 32% of follow-up antisocial behaviour was explained by the path from treatment through mediator to outcome. Additionally, a recently published paper (Hanisch et al., 2014) also conducted mediation analysis to assess whether the prevention program for externalizing problem behaviour (PEP) (Hanisch et al., 2010) improves child behaviour using the SEM approach. The findings suggested that changes in child externalizing problem behaviour were most strongly mediated by reductions of negative parenting in difficult parenting situations. Increases in positive parenting also served as a mediator, and changes in parental warmth, parents' feeling of self-efficacy, and parental mental health did not play a mediating role in the association between PEP treatment and child behaviour.

Listed in Table 1-2 are the intervention effectiveness results and the mediation analysis findings of a small number of well-designed parenting training RCTs that formally assessed

mediation. The table also includes a number of parenting training programme RCTs without formal mediation analysis but which tested the effect of the intervention on the potential mediators (Webster-Stratton et al., 2004, Webster-Stratton, 1984, Webster-Stratton, 1990, Webster-Stratton, 1994, Webster-Stratton et al., 1988, Webster-Stratton et al., 2001, Scott et al., 2010a, Scott et al., 2001b, Scott et al., 2012a, Sylva et al., 2008, Bagner and Eyberg, 2007, Markie-Dadds and Sanders, 2006, Scott et al., 2010b, Webster-Stratton and Hammond, 1997). Given the theories underlying these interventions, candidate mediators were most often measures of parent behaviour management skills, measures of general family functioning, or indices of youth association with deviant peers.

Table 1-2 Total Effects and Mediation Effects assessed in RCTs of Parenting Training Programmes for Child Behaviour Problems

Reference (RCT)	Sample Size	Intervention Type	Effectiveness Test	Candidate Mediators	Mediation Analysis Approach	Mediation Paths Assessment
Webster-Stratton, 1984	Mothers of 35 children with average age 4	VPT group VPT individual WL	Active treatments superior to WL	Parent commands Praise/criticism Spanking	NA	Not assessed VPT treatments had better effect on candidate mediators
Webster-Stratton et al., 1988	Parents of 114 children aged 3 to 8	VPT group VPT individual PT group WL	Active treatments superior to WL Trend toward VPT group superiority	Parent commands Praise/criticism Spanking	NA	Not formal mediation assessment Active treatments had better effect on candidate mediators than WL
Webster-Stratton, 1990	Parents of 43 children aged 3 to 8	VPT-T VPT-S WL	Active treatments superior to WL	Parent commands Praise/criticism Parental warmth Spanking	NA	No formal mediation assessment VPT treatments had better effect on spanking, praise, and warmth than WL
Webster-Stratton, 1994	Families of 78 children with average age 5	VPT VPT + CBT	No treatment superior	Child problem-solving Parent commands Praise/criticism Parent problem-solving Parental warmth Spanking Communication	NA	No formal mediation assessment VPT+CBT had better effect on parent and child problem-solving skills and marital communication Criticism, maternal stress, paternal commands, and paternal problem-solving predict deviant behaviour
Webster-Stratton & Hammond, 1997	Families of 97 children aged 4 to 8	IY PT IY CT IY PT+CT WL	Active treatments superior to WL	Child problem-solving Parent commands and criticisms Parental praise Parental positive affect Parental negative valence	NA	No formal mediation assessment CT and PT+CT had better effect on problem-solving and conflict management skills PT and PT+CT had better effect on positive interactions

				Spanking Couple collaboration Positive interaction		
Eddy & Chamberlain, 2000	Families of 79 youth aged 12-17	MTFC GC	MTFC treatment superior to GC	Family management skills Deviant peer association	SEM	Supervision, discipline, positive adult-youth relationship, and deviant peer association mediated the effect of treatment condition on youth antisocial behaviour
Scott et al., 2001b	Parents of 141 children aged 3 to 8	IY PTWL	IY treatments superior to WL	Praise Ineffective commands	NA	No formal mediation assessment IY had better effect on the ratio of praise and ineffective commands
Webster-Stratton et al., 2001	Parents and teachers of 272 children aged 4	IY PT+TTRS	IY treatments superior to RS	Negative and positive parenting styleParent–teacher bondingTeacher classroom management style	NA	No formal mediation assessment IY had better effect on all candidate mediators
Webster-Stratton et al., 2004	Parents and teachers of 159 children aged 4 to 8	IY PT IY PT+TT IY CT IY CT+TT IY CT+PT+TT WL	IY treatments superior to WL	Negative and positive parenting style Negative and positive teacher class room management style	NA	No formal mediation assessment IY had better effect on all candidate mediators
Gardner et al., 2006	Parents of 76 children aged 2 to 9	IY PT WL	IY treatment superior to WL	Observed positive parenting	SEM	Change in observed positive parenting mediated change in child negative behaviour
Markie-Dadds & Sanders, 2006	Families of 63 preschool-age children	Triple P WL	Triple P treatment superior to WL	Parenting style Parenting competence Parental adjustment	NA	No formal mediation assessment Triple P had better effect on use of dysfunctional discipline strategies and parenting competence

Bagner & Eyberg, 2007	Families of 30 children aged 3 to 6	PCIT WL	PCIT treatment superior to WL	Positive parenting Negative parenting Positive attention Effective commands Parenting stress	NA	No formal mediation assessment PCIT had better effect on increases in positive parenting, decreases in negative parenting, giving positive attention and using effective commands
Ogden & Hagen, 2008 and 1-year follow-up study Hagen *et al*., 2011	Parents of 112 children aged 4 to 12	PMTO RS	A significant main effect on observed total aversive behaviour for two-parent families.	Effective discipline Family cohesion	Regression models	PMTO predicted greater effective discipline and family cohesion at post-assessment, which in turn predicted improvements in several child domains at follow-up
Scott *et al*., 2010 and Sylva *et al*., 2008	Parents of 112 children aged 6	IY PT + child literacy Telephone helpline control	Active treatment superior to control on both child behaviour and child reading and writing skills	Parenting Style Parent use of reading strategies	Regression models for reading components	Not reading mediator is revealed IY had better effect on Praise Rewards, Time Out, Harsh discipline, Warmth, Criticism, Positive attention and Seek cooperation
Kling *et al*., 2010	Parents of 159 children aged 3 to 10	PMT-P PMT-S WL	PMT-P superior to PMT-S and WL at post-treatment and at 6-month follow up (PMT-P and PMT-S)	Harsh and inconsistent parenting (HI) Praise and incentives (PI), parent homework fidelity	Regression models + Sobel Test	Improvement in child conduct problems was mediated by improvement in parent competencies (-HI, +PI) and homework fidelity
Gardner *et al*., 2010	Parents of 153 children aged 2-8	IY Control	IY intervention/prevention superior to control	Positive parenting Negative parenting	Regression models + Sobel Test	Change in observed positive parenting mediated change in child negative behaviour

Scott *et al*., 2010a	Parents of 174 children aged 5 to 6	IY PT + shorter child literacy RS	No treatment superior	Child centred Negative control Positive affect Negative affect Calm discipline Praise/reward Coercive discipline	NA	No formal mediation assessment IY had better effect on child centred, negative affect and calm discipline
Scott *et al*., 2012a	Families of 215 children aged 5 to 7	IY Child literacy IY + child literacy Telephone helpline control	Active treatments superior to control on child behaviour IY superior to control on child reading	Positive Parenting Negative Parenting Parent use of reading strategies	NA	No formal mediation assessment IY had better effect on increased positive parenting (encouragement and praise) and decreased negative parenting (criticism, inconsistent discipline)
Hanisch et al., 2014	Parents of 155 children aged 3-6	PEP intervention Non-treated control	PEP intervention superior to non-treated control	Positive Parenting Negative Parenting parental warmth parental mental health	SEM	Changes in child externalizing problem behaviour were most strongly mediated by reductions of negative parenting in difficult parenting situations. Increases in positive parenting also served as a mediator. Changes in parental warmth, parents' feeling of self-efficacy, and parental mental health did not play a mediating role in the association between PEP treatment and child behaviour.

PMT-P: Practitioner-assisted Parent Management Training; PMT-S: Self-administrated Parent Management Training; WL: Wait-list control; IY PT CT TT: Incredible Years Parent Training Child Training Teacher Training; VPT: Videotape Parent Training; VPT-T: Videotape Parent Training plus therapist consultation; VPT-S: Videotape Parent Training self-administrated; CBT: Cognitive Behavioural Therapy; PTMO: Parent Management Training - The Oregon Model; RS: Regular Service; MTFC: Multidimensional treatment foster care; GC: Group Care control; PCIT: Parent–Child Interaction Therapy; Triple P: Positive Parenting Program; PEP: Prevention program for preschool children with Externalizing Problem behaviour

1.4 Statistical methodology for mediation analysis

This section introduces general mediation and moderation related terminologies, defines causal mediation parameters of interest, provides a review of standard mediation analysis and summarizes outstanding statistical challenges.

1.4.1 A brief introduction to mediators, moderators, confounders and predictors

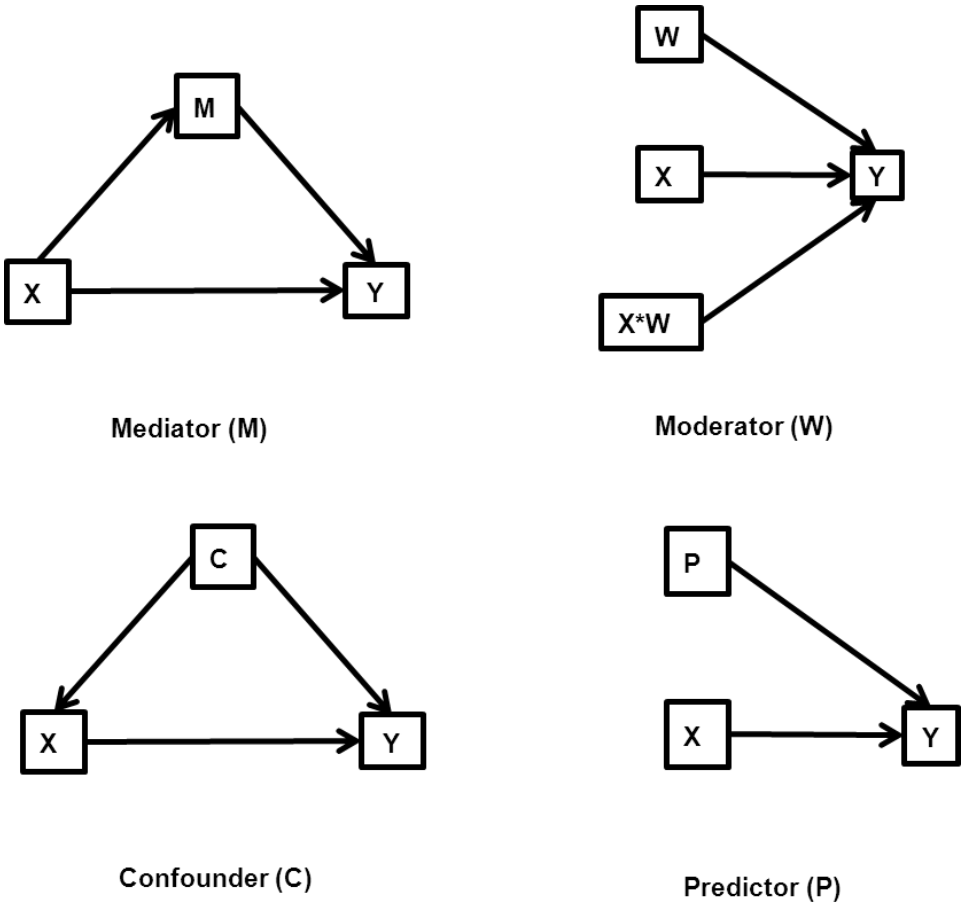
When we study the causal relationship between a dependent variable *Y* and an independent variable *X*, it is possible that there is a third variable *M* that is an intermediate variable in the causal chain such that *X* causes *M* and *M* causes *Y*. The variable *M* is called a *mediator* and this type of relationship is called mediation (MacKinnon, 2008). For example, exposure to negative life events affects blood pressure through the mediation of cognitive attributions to stress. Mediation is different, however, from moderation. According to Baron and Kenny (Baron and Kenny, 1986), "a moderator is a qualitative (e.g. sex, race, class) or quantitative (e.g. level of reward) variable that affects the direction and/or strength of the relation between an independent variable and a dependent or criterion variable". Moderator effects are typically known as interactions *X*W* with the effect of an independent variable *X* on a dependent variable *Y* depending on the levels of the *moderator W*. For example, the effects of a behavioural therapy may be much greater for men than for women, so that gender is the moderator of the therapy effects. When *X* is a treatment group, moderators are also typically referred to as *treatment effect modifiers* in the medical literature (A'Campo et al., 2012). Furthermore, if the baseline variable *W* is a biological marker modifying the treatment effect, then it is also a *predictive marker* (Simon, 2008).

Another situation in which the observed association between a dependent variable *Y* and an independent variable *X* may be altered is when a third variable *C* is a common cause of both *X* and *Y*. Such a *C* variable is a *confounder* of the effect of *X* on *Y*. A confounder is defined as a variable *C* that causes both the independent variable *X* and the dependent variable *Y* and falsely obscures or accentuates the relationship between *X* and *Y* (Greenland and Morgenstern, 2001). For example, experiment results may show negative relationship between bottle-feeding and diarrhoea in infants. However, it would seem logical that bottle-fed infants are more prone to diarrhoea since water and the bottle could get contaminated, milk could go bad, etc. This experimental negative relationship might be due to the confounding effect of mother's education as the experiment findings might be that better-

educated mothers are more likely to bottle-feed their infants, who are also less likely to develop diarrhoea due to better hygienic practices of the mothers.

Different from the situations above, it is possible that both an independent variable of interest *X* and another independently distributed variable *P* cause a dependent variable *Y*. In this case, both *X* and *P* are *predictors* of *Y*. The predictor *P* will not change the observed relationship between *X* and *Y*, but will make the prediction of *Y* more accurate because it explains variability in the *Y* variable (MacKinnon, 2008). For example, age may predict the outcome of a child's writing skills for both a randomised intervention and a control group at the same level. In the clinical literature, a baseline variable that predicts the outcome variable within each trial arm would be called a *prognostic marker* (Simon and Altman, 1994). (Note that a predictor by definition is neither a confounder nor a mediator of the *X*-*Y* relationship). The three graphs in Figure 1-1 illustrate these definitions.

Figure 1-1: Path diagrams illustrating the definitions of mediators, moderators, confounders and predictors (directed paths imply causal effects)



Both confounders and mediators account for the relationship between *X* and *Y*, but mediators stand as part of a causal mediation process. The mediator explains the relation between *X* and *Y* because it transmits the effect of *X* on *Y* through the mediator *M* (MacKinnon et al., 2007a). The concept of a moderator is quite different from that of a confounder: The confounder distorts the observed association and is a factor one hopes to prevent or control for when investigating the true relationship between the dependent and independent variables (Weinberg, 1993). In contrast, the moderator variables interact with independent variable *X*, so that the relation between *X* and *Y* is different at different levels of the moderator variable. Thus effect moderation is a more detailed description of the effect itself (Aguinis, 2004). In addition to the detailed discussion of the distinction between mediators and moderators in Baron and Kenny's famous paper (Baron and Kenny, 1986), it is specified in Karemer et al.'s paper (Kraemer et al., 2001) that a useful characteristic distinguishing mediators from moderators is that mediators change in response to an intervention, whereas moderators are measured at baseline before receiving an intervention. In a trial, both moderators and predictors of treatment can be baseline characteristics. While treatment effect moderators differentially predict outcome across treatment groups, predictors predict outcome regardless of treatment condition (Beauchaine et al., 2005).

1.4.2 Causal inference mediation framework

The concept of mediation concerns the extent to which the effect of one variable on another is mediated by some possible intermediate variable, so that mediation concerns causality. There has been a trend in recent decades to embed the emerging literature on causal mediation analysis within the causal inference framework. For the understanding of causal mediation, it is necessary to formally define a causal effect. Even though we all have an intuitional understanding of a causal effect, the mathematical notation can formalize this causal intuition and provide a precise definition (Hernán and Robins, 2015).

Let us start with the definition of an *Individual Treatment Effect* (ITE). Consider a dichotomous intervention variable R_i ($R_i = 1$ for treated; $R_i = 0$ for not treated), and an observed outcome variable Y_i that can be any value for the i-th subject (observational unit). Let $Y_i(0)$ be the variable that would have been observed under the treatment value $R_i = 0$

and let $Y_i(1)$ be the variable that would have been observed under the treatment value $R_i = 1$. This leads to the definition of the ITE as the contrast

$$ITE = Y_i(1) - Y_i(0) \qquad \text{Equation 1-1}$$

with treatment R having a causal effect on the individual's outcome Y_i if $Y_i(0) \neq Y_i(1)$. $Y_i(0)$ and $Y_i(1)$ are referred to as *potential outcomes* or *counterfactual outcomes*. The former emphasizes that depending on the treatment that is received, either of these two outcomes can potentially be observed. The latter emphasizes that these outcomes represent situations that may not actually occur (that is, they are counter to the fact). For each subject, only one of the counterfactual outcomes is actually observed (factual), i.e. if a subject actually received treatment $R_i = r$, the observed outcome Y_i equals this counterfactual outcome $Y_i(r)$.

The fundamental problem of causal inference is that identifying individual causal effects from observed data is generally not possible. Thus we turn our attention to the *Average Treatment Effect* (ATE) in a population of individuals. Let $E[Y(0)]$ be the mean outcome if all subjects in the population received treatment $R = 0$ and let $E[Y(1)]$ be the mean outcome if they all received $R = 1$, where 'E'xpectation refers to the population mean or average. Then ATE is defined as the mean ITE:

$$ATE = E[Y(1)] - E[Y(0)] = E[Y(1) - Y(0)] \qquad \text{Equation 1-2}$$

To identify a total causal treatment effect (ATE), we need to assume that $Y(r)$ is independent of R, $Y(r) \perp R$. This is referred to as the no-confounders assumption. This project is concerned with estimating treatment effects from trials with *R* denoting randomly allocated treatment ($R = 1$ "treatment offered" and $R = 0$ "treatment not offered"). It is assumed throughout that those offered treatment also receive it, and similarly that those who are not offered it do not receive it; thus, causal treatment effects are here defined as contrasts between the two counterfactual situations $R = 1$ and $R = 0$.

Having defined a total causal treatment effect (ATE), I now introduce formal definitions of mediation parameters. It begins by defining a list of mediator M related potential (counterfactual) outcomes:

$M(0)$: Mediator if randomised to control group. This is the intermediate outcome that would have been observed under the treatment value $R = 0$.

$M(1)$: Mediator if randomised to treated group. This is the intermediate outcome that would have been observed under the treatment value $R = 1$.

$Y(r, m)$: Outcome with treatment r and mediator m. Here, the counterfactual concept is extended to treatment-mediator joint exposure (R, M). $Y(r, m)$ is the outcome that we would (possibly contrary to fact) have observed for the treatment had R been set to the value r and, likewise, M to the value m, through some intervention or manipulation.

$Y(0) = Y(0, M(0))$: Outcome if randomised to control group with intermediate outcome $M(0)$. Compared to $Y(r, m)$, the key difference is that $M(0)$ is a counterfactual intermediate outcome instead of a manipulated value m.

$Y(1) = Y(1, M(1))$: Outcome if randomised to treated group with intermediate outcome $M(1)$. In short, $Y(1)$ is the outcome that would have been observed under the treatment value $R = 1$, while the mediator also takes the would-be value under the treatment value $R = 1$.

In the control arm, $Y(0)$ and $M(0)$ are observed, $Y(1)$ and $M(1)$ are missing. In the treated arm, $Y(1)$ and $M(1)$ are observed, $Y(0)$ and $M(0)$ are missing. $Y(r, m)$, $Y(0, M(0))$, and $Y(1, M(1))$ are the counterfactual concepts defined by extending to the joint exposure (R, M) where M is the potential mediator, then the causal direct and indirect effects are defined based on these extended counterfactual concepts (Pearl, 2001, Robins and Greenland, 1992). Given the differences of nomenclature, a summary of the definitions of direct and indirect effects has been conducted by VanderWeele and Vansteelandt (VanderWeele and Vansteelandt, 2009).

Assuming that we can physically set the level of the mediator to a specific value, for a dichotomous intervention variable R ($r = 1$ for treated; $r = 0$ for not treated) the *Controlled Direct Effect* of intervention R on the outcome Y for a fixed level of the mediator at m is defined as

$$CDE = E[Y(1, m) - Y(0, m)] \qquad \text{Equation 1-3}$$

However, it is not realistic to assume that the mediator can be forced to be the same for all subjects in the population under most scenarios. Actually, the indirect effect cannot be defined in the same way as the controlled direct effect, as it is impossible to hold a set of variables fixed in such a way that the effect of exposure on outcome would circumvent the direct pathway. In particular, the difference between the total effect and the controlled direct effect, $E[Y(1) - Y(0)]$ minus $E[Y(1, m) - Y(0, m)]$ may not be interpreted as an indirect effect.

To overcome the limitations above, natural direct effects and indirect effects are defined (Pearl, 2001, Robins and Greenland, 1992). Instead of controlling the mediator at a fixed level, the natural direct and indirect effects consider if the mediator were kept at the level it would have taken under the corresponding intervention.

The *Natural Direct Effect* is defined as

$$NDE = E[Y(1, M(0)) - Y(0, M(0))] \qquad \text{Equation 1-4}$$

It expresses the effect that would be realized if the intervention were administered but its effect on the mediator were somehow blocked, or equivalently, if the mediator were kept at the level it would have taken in the absence of the intervention. This quantity is also referred as the *Pure Direct Effect (PDE)* (Robins, 2003).

The *Natural Indirect Effect* is defined as

$$NIE = E[Y(1, M(1)) - Y(1, M(0))] \qquad \text{Equation 1-5}$$

It expresses how much the outcome would change on average if the intervention were kept as treated, but the mediator were changed from level M(0) to M(1). This quantity is also referred as the *Total Indirect Effect (TIE)* (Robins, 2003).

The *Pure Indirect Effect (PIE)* defined by Robin and Greenland (Robins and Greenland, 1992) is $PIE = E[Y(0, M(1)) - Y(0, M(0))]$.

The sum of *Natural Direct Effect* and *Natural Indirect Effect* equals the *Average Treatment Effect* (total causal treatment effect):

$$NDE + NIE = ATE \qquad \text{Equation 1-6}$$

Identification of the direct and indirect effects requires some or all of the assumptions below:

(1) no confounders of the association between treatment and outcome that $Y(r, m) \perp R$ for all levels of r and m;
(2) no confounders of the association between mediator and outcome that $Y(r, m) \perp M$ for all levels of r and m;
(3) no confounders of the association between treatment and mediator that $M(r) \perp R$ for all levels of r;
(4) no interaction assumption that $Y\ (r, m) - Y\ (0, m) \perp M$ for all levels of r and m.

In brief, the identification of ATE requires assumption (1); the identification of CDE requires assumptions (1) and (2); the identification of NDE requires assumptions (1), (2) and (3); and the identification of NIE requires all four assumptions. Specifically, with assumption (4), the direct effect is independent of the levels of the mediator; the *Controlled Direct Effect* equals the *Natural Direct Effect*:

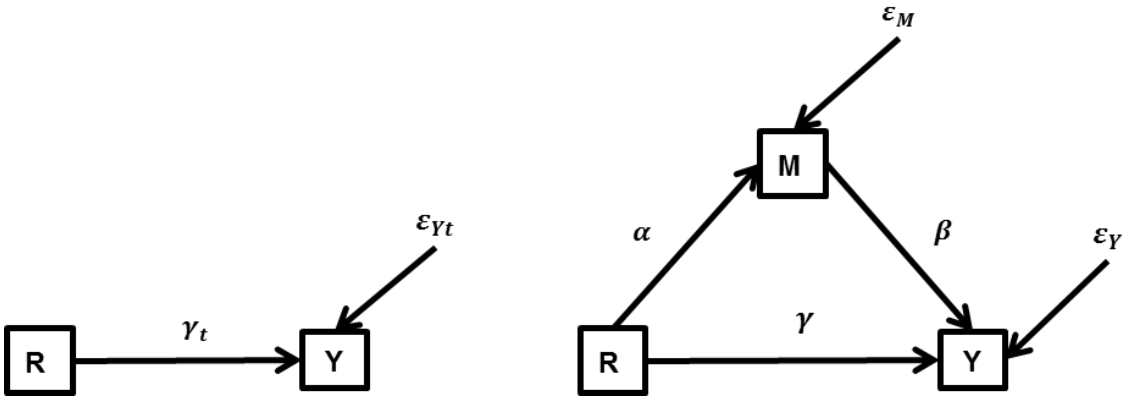
$$CDE = NDE \qquad \text{Equation 1-7}$$

Having introduced the concepts of total causal effect, direct and indirect effects and their identification assumptions, the next section will review the methods of standard mediation analysis, with the emphasis on showing how to use regression to estimate the causal mediation effects of interest (the total, direct and indirect effects).

1.4.3 Review of standard mediation analysis approaches

Standard approaches to mediation analysis commonly used in the field of social and behavioural sciences include regression and structural equation modelling (SEM) (Ten Have and Joffe, 2010). The use of the standard mediation approaches in psychology has drawn upon the widely cited work of Baron and Kenny (Baron and Kenny, 1986), and the methodological work of Mackinnon (MacKinnon, 2008) has also been highly influential in this area. Figure 1-2 shows the single-mediator model and three association equations are also illustrated as follows (Mackinnon and Dwyer, 1993):

Figure 1-2: The single-mediator mediation model diagram for RCT



Standard mediation approaches are based on three regression equations for continuous *Y* and continuous *M* respectively:

$$Y = i_0 + \gamma_t R + \varepsilon_{Yt} \qquad \text{Equation 1-8}$$

$$Y = i_1 + \gamma R + \beta M + \varepsilon_Y \qquad \text{Equation 1-9}$$

$$M = i_2 + \alpha R + \varepsilon_M \qquad \text{Equation 1-10}$$

where R denotes randomly allocated treatment, M denotes the mediator and Y denotes the dependent variable. The coefficient γ_t represents how strongly R predicts Y; γ is the coefficient representing the strength of prediction of Y from R when holding M at a fixed level; β represents the strength of the relationship between M and Y within fixed levels of R; and α is the coefficient representing the strength of the relationship between R and M. The intercepts in each equation represent the value of each variable if $R = 0$ $(and\ M = 0)$

are i_0, i_1 and i_2 respectively; ε_{Yt}, ε_Y and ε_M represent the error, or the part of the relationship that cannot be predicted.

According to Baron and Kenny (Baron and Kenny, 1986) and Judd and Kenny (Judd and Kenny, 1981), the classic mediation regression approach involves estimating three regression models: the parameters in Equation 1-8, Equation 1-9 and Equation 1-10. These three regression equations provide three separate tests: significance tests of the strength of the total (overall) relationship between R and Y, (γ_t), the strength of the relationship between R and M, (α), the strength of the relationship between M and Y adjusted for R, (β), and a visual inspection of whether $\widehat{\gamma_t}$ (the hat ^ symbol here denotes an estimator of the respective regression coefficient) is greater than $\hat{\gamma}$. Perfect mediation holds if the independent variable has no effect when the mediator is controlled for; in these circumstances, γ drops to zero. Despite the wide use of Baron and Kenny's (1986) approach, the requirement of the significant overall relation between R and Y has been criticised by some researchers (MacKinnon, 2008) who claimed that mediation can exist even in the absence of such an overall significant relationship.

The Baron and Kenny approach estimates the product $\alpha\beta$ and interprets this as "mediated effect". This product term can be estimated from a sample using Ordinary Least Squares (OLS) regression in one of two ways: Using OLS estimates of $\hat{\alpha}$ and $\hat{\beta}$ to construct $\hat{\alpha}\hat{\beta}$ or using OLS estimates of $\widehat{\gamma_t}$ and $\hat{\gamma}$ to construct $\widehat{\gamma_t} - \hat{\gamma}$ (Mackinnon and Dwyer, 1993). Under regression equations Equation 1-8, Equation 1-9 and Equation 1-10, the estimators $\widehat{\gamma_t} - \hat{\gamma}$ and $\hat{\alpha}\hat{\beta}$ always give the same value, unless different samples were used to estimate the regression equations. For example, the sample size for Equation 1-8 might be different from the sample size used to estimate Equation 1-9 and Equation 1-10, if some subjects have missing observations of the mediator variable.

The estimate of the mediated effect and its standard error can be used to construct confidence intervals (CI = mediated effect $\pm$ $z_{1-\alpha/2}$ × standard error), where $z_{1-\alpha/2}$ is the value of z (or t) statistics for the required confidence limits (e.g. 1.96 for 95% confidence limits for a large sample size). The confidence limits are the upper and lower $(\alpha/2)$ quantile of cumulative distribution function for 1- α confidence level. Although the α symbol is used to denote both the significance level and the relationship parameter between *R* and *M*, it

should not be problematic to distinguish them in context. The standard error of $\hat{\alpha}\hat{\beta}$ is referred to as the *product of coefficient standard error*, and the standard error of $\hat{\gamma}_t - \hat{\gamma}$ as the *difference in coefficients standard error*.

The most commonly used formula for the standard error of $\hat{\alpha}\hat{\beta}$ was derived by Sobel (1982). Equation 1-11 of the standard error $S_{\hat{\alpha}\hat{\beta}}$ is shown below, where $S_{\hat{\alpha}}$ is the standard error of $\hat{\alpha}$ and $S_{\hat{\beta}}$ is the standard error of $\hat{\beta}$.

$$S_{\hat{\alpha}\hat{\beta}} = \sqrt{\hat{\alpha}^2 s_{\hat{\beta}}^2 + \hat{\beta}^2 s_{\hat{\alpha}}^2} \qquad \text{Equation 1-11}$$

The derivation of Sobel standard error in Equation 1-11 is accomplished using covariance algebra based on Equation 1-9 and Equation 1-10 and the assumption of uncorrelated error terms across equations, normally distributed variables, and a linear system of relations among variables. Thus, use of the estimator of the standard error in Equation 1-11 $S_{\hat{\alpha}\hat{\beta}}$ assumes independence between $\hat{\alpha}$ and $\hat{\beta}$. If there is a non-zero covariance between $\hat{\alpha}$ and $\hat{\beta}$, for example, in some structural equation models (such as latent variable mediation models: see the following section), then this formula must be expanded.

Although the Sobel standard error estimate of the indirect effect may be unbiased, there is evidence that confidence interval based on these values do not perform well (MacKinnon et al., 2007b). Two extensive simulation studies (MacKinnon et al., 1995, Stone and Sobel, 1990) showed an imbalance in the number of times a true value fell to the left or right of the confidence interval. The use of the standard error of $\hat{\alpha}\hat{\beta}$ in Equation 1-11 to construct confidence intervals assumes that the product $\hat{\alpha}\hat{\beta}$ has a normal distribution. In fact, the distribution of the product of two random variables does not necessarily have a normal distribution, especially when the sample size is small. In order to relax this normality assumption, resampling based confidence intervals approaches have been developed and are recommended for standard mediation analyses. I will return to this issue in Chapter 3 and describe in detail a non-parametric bootstrap approach for constructing confidence intervals and significance tests for mediation effects.

Mackinnon et al. (MacKinnon et al., 2002) provide Equation 1-12 for the standard error of $\widehat{\gamma_t} - \hat{\gamma}$, where $rs_{\widehat{\gamma_t}}s_{\hat{\gamma}}$ is the covariance between $\widehat{\gamma_t}$ and $\hat{\gamma}$.

$$S_{\widehat{\gamma_t}-\hat{\gamma}} = \sqrt{s^2_{\widehat{\gamma_t}} + s^2_{\hat{\gamma}} - 2rs_{\widehat{\gamma_t}}s_{\hat{\gamma}}} \qquad \text{Equation 1-12}$$

The values from Equation 1-11 and Equation 1-12 are very similar. However, the computation and generalization of Equation 1-12 are more complicated than Equation 1-11, so that the former is typically used.

A structural equation modelling (SEM) approach combines a measurement structure (measurement model) with the structural model, in which the latent (or unobserved) constructs are formed by separating true and error variance in observed measures (Bentler and Dudgeon, 1996). SEM (including path analysis) specifies the variables' relationships in matrix form, from which a parameterisation of the covariance matrix is obtained. The parameters describing the covariance matrix are estimated by moment methods or maximum likelihood (ML), and corresponding standard errors can be obtained using SEM computer programs such as LISREL (Jöreskog and Sörbom, 1996 - 2001), Mplus (Muthén and Muthén, 1998-2010), EQS (Bentler and Wu, 2005) and AMOS (Arbuckle, 1995–2007). When putative mediator and outcome constructs have been observed, estimates and associated inferences for the product term $\alpha\beta$ obtained by fitting a structural equation model (by ML or methods of moments) are very similar to OLS-based inferences from separate regression models, and results are identical for large sample sizes.

Recent contributions in mediation analysis have emphasized the importance of articulating identifiability conditions for a causal interpretation. The notations of direct and indirect causal effects from causal inference in the counterfactual framework provide a clearer causal interpretation (see Section 1.4.2). It is shown that concepts of direct and indirect effect from causal inference generalize those described by Baron and Kenny. Linking $\alpha\beta$ and γ to the counterfactual causal parameters of interest (the causal direct and indirect effects), $\alpha\beta$ can be conceived as NIE, γ can be conceived as NDE and the sum of $\alpha\beta$ and γ is ATE(total causal treatment effect). Expressed in equations, this gives:

$$NIE = \alpha\beta \quad \text{Equation 1-13}$$

$$NDE = \gamma \quad \text{Equation 1-14}$$

$$ATE = \alpha\beta + \gamma \quad \text{Equation 1-15}$$

For these equalities to hold, linear regression approaches in RCTs need to make identifiability assumptions, including:

(1) linearity in the regression equations of continuous outcome $Y$and continuous mediator M;
(2) no confounding (observed or unobserved) between mediator $M$and outcome Y;
(3) no treatment mediator ($R \times M$) interactions.

As shown in Equation 1-7, $CDE = NDE$ under the assumption of no $R \times M$ interaction. In this case,

$$NDE = CDE = \gamma \quad \text{Equation 1-16}$$

For the simplicity of notations, in this project I express that $\alpha\beta$ is the *Indirect Effect* (IE), γ is the *Direct Effect* (DE), and $\alpha\beta + \gamma$ is the *Total Effect* (TE).

The counterfactual definitions and results for direct and indirect effect have also been extended when nonlinearities are present (Robins and Greenland, 1992, VanderWeele and Vansteelandt, 2009). Under appropriate identification assumptions, these more general direct and indirect effects from causal inference can be estimated using regression even when there are interactions between the primary exposure of interest and the mediator. In this PhD project, I still make the assumptions of linearity and no treatment mediator interaction ($R \times M$) while the development of the statistical methodology in mediation analysis of RCTs will focus on addressing the statistical challenges listed in the following section.

1.4.4 Statistical challenges

Traditional mediation analysis approaches such as the regression approach (Baron and Kenny, 1986) assume the absence of mediator-outcome confounding in the regression models in order to estimate the causal parameters of interest. Using RCTs to investigate mediation, one can obtain the independence between treatment and outcome ($Y \perp R$) and the independence between treatment and mediator ($M \perp R$). However, the Baron and Kenny approach in RCTs still needs to assume that there is no confounding (observed or

unobserved) between mediator M and outcome Y (see Section 1.4.3, assumption 2). This assumption might be overly simplistic in some RCTs. Since both M and Y are outcomes (one is an intermediate outcome, while the other is a distal outcome), one cannot rule out the existence of variables which affect both M and Y. Therefore, the mediation parameters estimated by traditional mediation analysis might be subject to confounding bias.

Missing values are likely to be present in observed confounding variables, putative mediator and clinical outcome variables. A popular approach (and default approach in many statistical software packages) used to deal with missing values is to restrict the sample to complete cases only: in our context, this means that only study participants who provide observations for the parenting mediator, the child outcome and the confounders are included in the analysis (known as *complete case analysis*, CC analysis). Such an approach is inefficient, as it discards some of the information available. It also relies on restrictive assumptions regarding the process that has generated the missing values: Specifically, CC analysis assumes that the missing values' patterns are only predicted by variables that feature as explanatory variables in the respective regression model (White and Carlin, 2010). Wrong assumptions regarding the process that generates the missing data lead to bias in estimators of population parameters of interest. As reviewed in Section 1.4.3, traditional mediation analysis is in fact based on two linear regression models, so estimators of causal mediation parameters derived by CC analyses might suffer such missing data biases.

Traditional linear regression mediation analysis assumes that individual observations are independent of each other. However, interventions may be delivered in groups with shared experiences (e.g. the same therapist). Thus it is possible that observations from participants from the same group might be correlated. Additionally, the experimental design can also yield clustered data: For example, cluster randomised trials in which a group of participants are randomised to the same intervention (the randomisation unit is a cluster of participants). Such clusters share characteristics which again may lead to more similar observations for participants from the same cluster. This correlation is so-called *intra-cluster correlation* (ICC). The existence of intra-cluster correlation violates the assumption of independent observations in the traditional linear regression mediation analysis. Failure to take clustering into account is likely to lead to biased estimators of mediation effects.

It is possible that the scale of the mediator (and outcome) is discrete instead of continuous. Although the linearity assumption might still be acceptable, it might not be realistic to assume normality of the distributions of the mediator (and outcome) to be investigated. Implausible assumption of the variable distribution may lead to biased inference to the estimation of mediation effects. In addition, a standard method to construct the confidence interval for an indirect effect in traditional mediation analysis uses the Sobel standard error of $\hat{\alpha}\hat{\beta}$ in Equation 1-11. The confidence interval based on Sobel standard error assumes that the product $\hat{\alpha}\hat{\beta}$ has a normal distribution. In fact, the distribution of the product of two random variables is not normal in most scenarios (MacKinnon and Fritz, 2007). Thus it might be preferable to base inferences on a non-parametric method that does not rely on a normality assumption.

Power and sample size issues are often a major limitation of mediation analysis, as approaches require large sample sizes to estimate relevant parameters with adequate precision. Simulation results (Fritz & Mackinnon, 2010) provide that the empirical estimates of sample sizes needed for 0.8 statistical power using the Sobel test are: 667 when both α and β are small ($\alpha = 0.14, \beta = 0.14$), 422 when α has medium size and β is small ($\alpha = 0.39, \beta = 0.14$), and 90 when both α and β have medium size ($\alpha = 0.39, \beta = 0.39$). It seems that for small/medium size of α and β, several hundreds of samples are required. In reality, the sample size of a single RCT may not meet the requirement for achieving 0.8 statistical power to detect a mediation effect. Pooling data from multiple trials for the purpose of meta-mediation analysis may help solve these issues. However, this requires identification and application of clear inclusion criteria to define target populations and appropriate measurement of concepts. I will discuss meta-mediation analysis of pooled data further in Chapter 5.

1.5 Thesis outline

In the following chapters of my PhD thesis, I describe in detail my statistical methods of mediation analysis for trials of parenting intervention allowing unmeasured confounding in the presence of missing data and report the mediation analysis results using the example of three trials of IY parenting programmes. Chapter 2 begins by introducing three RCTs of IY parenting programmes and records the trials' data harmonisation for the combined data mediation analysis. Chapter 3 is concerned with developing a new mediation analysis

method (MI-BT) that accounts for measured confounding of the mediator-outcome relationship in the presence of missing data. To handle missing data I used a new inferential approach that makes use of all the available data and has less restrictive assumptions regarding the missing data generating process than CC analysis. The approach I developed facilitates the use of linear mixed models to reflect aspects of trial design (e.g. cluster randomisation, group treatments), employs multiple-imputation by chained equation (MICE) to construct consistent estimators of mediation parameters, and generates non-parametric inferences via a cluster bootstrap (BT) approach. Application of this MI-BT method to the first trial analysed, SPOKES, showed statistically significant indirect effects for two variables parental criticism and parental warmth. Chapter 4 describes how I extend the method for the trials of parenting intervention to relax the no unmeasured confounding assumption using the Instrumental Variable (IV) approach, which becomes the IV-MI-BT approach. Finding variables that can act as instruments is crucial. To address this I suggested a strategy for constructing a list of potential instrumental variables and selecting the most promising ones. I found a number of IVs for parental criticism (interactions between randomisation and baseline parental characteristics such as depression and level of education, therapy groups), and a different IV for parental warmth (number of therapy sessions). Application of this IV-MI-BT method to the SPOKES trial showed that while IV estimators of causal mediation parameters were similar in value compared to previous estimates, their confidence intervals were inflated. Chapter 5 introduces the procedure for conducting an IV meta-mediation analysis using individual participant's data (IPD) from three parenting trials for the purpose of regaining precision. A framework for conducting such a meta-mediation analysis was developed. This includes systematic steps explaining how to implement: (i) different trial designs in the combined analysis model, (ii) a parameterisation that enables empirical assessment of effect heterogeneity across trials, (iii) MI-BT-based method for testing effect heterogeneity (interactions) and (iv) inferences for direct and indirect parenting programme effects based on the final model. Meta-analysis of the three contributing trials did not detect any evidence for between-trial heterogeneity in mediation effects. Pooling of the studies resulted in smaller and non-significant overall indirect effect estimates and provided a considerable precision gain compared to the SPOKES-only analysis. Finally, the discussion chapter reviews both the novelty and the limitation of the statistical methods developed in this project and recommends possible areas for further research.

Chapter 2 Trials of Three Parenting Programmes

This project applies mediation analyses to data from three trials of parenting programmes conducted by researchers from the Institute of Psychiatry, King's College London. They are the Supporting Parents On Kids' Education in Schools (SPOKES) trial, the Clinical Parenting Trial (CPT) and the Help Children Achieve (HCA) trial. All three trials collected measures of the distal child outcome (child antisocial behaviour) as well as measures of the intermediate outcome (aspects of parenting) targeted by the respective interventions in order to improve child outcome.

SPOKES (Scott et al., 2010b) was a randomised controlled trial of a parenting group intervention for improving child antisocial behaviour in eight schools in London. This succeeded the earlier waiting list controlled trial, CPT (Scott et al., 2001b), of a parenting group intervention for childhood antisocial behaviour in clinical practice. Finally, the most recent trial, HCA (Scott et al., 2012b) was a factorial randomised controlled trial of two types of parenting groups (including separate behavioural and literacy programmes). A brief summary of the three parenting trials is presented in Table 2-1.

In the following three sections, I introduce the three trials separately in terms of target population, trial design, components of the intervention under study and outcome measurements. The SPOKES trial will be used as an example to demonstrate the mediation analysis approaches proposed in Chapters 3 and 4. The SPOKES, CPT and HCA trials will be pooled together to form a combined data set for the meta-mediation analysis in Chapter 5. However, different trials might apply different measurement instruments and different scales to measure the confounders, the mediators and the outcome. This makes it impossible to combine the data from the three trials directly. Thus, approaches to achieve compatible measures of the same concept are required for pooling the trials together, and I refer to these approaches as *harmonisation approaches* that will be applied to combine the three parenting trials are detailed in Section 2.4. Finally, the supporting information for the existence of trial effect and a comparison between the harmonisation approaches and the complex modelling approach are discussed in Section 2.5.

Table 2-1 Summary Table of Three Trials of Parenting Programmes

Aspects	Trial Name		
	SPOKES	**CPT**	**HCA**
Trial Design	Randomised Controlled Trial – two arms	Waiting List Controlled Study – two arms	Randomised Controlled Trial – four arms
Trial period	1999 – 2001	1995 – 1999	2008 – 2012
Treatment Groups	Intervention group: IY + Literacy (22 sessions) Control group: Telephone helpline	Intervention group: IY (12 sessions) Control group: Waiting list	Intervention group 1: IY + Literacy (22 sessions) Intervention group 2: IY (12 sessions) Intervention group 3: Literacy (10 sessions) Control group: Telephone helpline
Participants	Parents of 5-6-year-old children who have high antisocial behaviour scores in 8 schools in Lambeth, London, among the 5% most deprived English Boroughs.	Parents of children aged 3 – 8 years referred to four NHS child and adolescent mental health services because of antisocial behaviour.	Parents of children aged 5 – 7 who have high antisocial behaviour scores in a disadvantaged inner London Borough and a South West city.
Sample Size	Total: n=112 Intervention group: n=61 Control group: n=51	Total: n=110 Intervention group: n=73 Control group: n=37	Total: n=213 Intervention group 1: n=50 Intervention group 2: n=56 Intervention group 3: n=53 Control group: n=54
Primary Child Outcome Measurements	Child Antisocial Behaviour measured using PACS semi-structured interview at 1 year after randomization	Child Antisocial Behaviour measured using PACS semi-structured interview at 5 – 7 months after randomization	Child Antisocial Behaviour measured using PACS semi-structured interview at 9 – 11 months after randomization
Parent Outcome Measurements	Positive Parenting Practices and Negative Parenting Practices measured using interview, questionnaire and direct observation		

IY: "Incredible Years" parent behavioural training programme
PACS: Parent Account of Child Symptoms

2.1 Overview of SPOKES trial

2.1.1 Participants

The trial ran from 1999 to 2001 in eight schools in Lambeth, London, among the 5% most deprived English Boroughs. All children in Reception and Year One classes were screened. Both teachers and parents were asked to complete the Conduct Problems scale of the SDQ (five questions) and the DSM-IV (eight questions) and the scores from parents and teachers were summed. The cut-off was the summed score SDQ $\geq$ 5 or DSM $\geq$ 10, one standard deviation above the population mean for 5-6-year-olds, designed to capture most cases at risk of lifetime-persistent antisocial behaviour. Eligible children had to exhibit conduct

symptoms above the screen cut-off level. Additional inclusion and exclusion criteria were: (1) ability to understand English; (2) ability to attend at group times; (3) interest in attending; (4) willingness to participate in a randomised trial; (5) child free of clinically apparent developmental delay. At the end, the parents of 112 5-6-year-old children who had high antisocial behaviour scores (above the cut-off) were included in the trial.

2.1.2 Trial design

SPOKES was a randomised controlled trial for evaluating the population-based parenting intervention that was aimed to tackle early-onset antisocial behaviour. The eight schools in the three years during which the study recruited formed ten strata within which individual parents were randomly allocated to the intervention or the control group. In total, sixty-one parents were randomised to the intervention group and fifty-one parents were randomised to the control group. Table 2-2 lists the counts of participants in each stratum for the control and the intervention group.

Table 2-2 Number of participants in each stratum by trial arm for SPOKES

School-year Stratum	No. Control	No. Intervention	Total
I	4	4	8
II	3	4	7
III	2	7	9
IV	4	4	8
V	6	11	17
VI	7	5	12
VII	4	6	10
VIII	3	6	9
IX	5	9	14
X	13	5	18
Total	51	61	112

2.1.3 Intervention

The parenting intervention was delivered in schools to groups of four to eight parents for two hours one morning per week. The children were not seen. The SPOKES trial intervention provided a twelve-week behavioural programme followed by a ten-week literacy programme. The twelve-week “Incredible Years” parent behavioural training programme (Webster-Stratton et al., 2008) included videotape clips of parents with their children. The content covered promotion of desirable child behaviour and on-task attending through play,

praise and rewards, handling misbehaviour, applying consequences, and time out. Through detailed group discussion and role play, the parental behaviour that leads to better child behaviour was drawn out and practised. The child literacy programme was a manualised programme (Sylva et al., 2008) that trains parents to help school-age children to increase their ability to independently read texts of an appropriate level of difficulty.

Parents randomised to the control group were offered a telephone helpline manned by the same staff, who advised them how best to access regular services. This intervention had the advantage of being brief and flexible.

Table 2-3 Counts of participants in each therapy group by school strata for SPOKES trial

Therapy Group	School-year Stratum										
	I	II	III	IV	V	VI	VII	VIII	IX	X	Total
A	4										4
B		4									4
C			7								7
D				4							4
E					3						3
F					8						8
G							6				6
H								6			6
I									9		9
J						5					5
K										5	5
Intervention	4	4	7	4	11	5	6	6	9	5	61
Control	4	3	2	4	6	7	4	3	5	13	51
Total	8	7	9	8	17	12	10	9	14	18	112

By design, the therapy groups were nested within school-year strata in the intervention arm, i.e. one or more therapy groups were run for each stratum. Table 2-3 presents the numbers of participants in each therapy group by school-year strata. This implies a hierarchical structure of the SPOKES data: In the intervention arm, it has a three-level hierarchical structure with parents (level 1) nested within therapy groups (level 2) that were nested within school-year strata (level 3); in the control arm, it has a two-level hierarchical structure with parents (level 1) nested within school-year strata (level 2). Since parent training was delivered in groups, the outcomes from the same therapy group would be expected to be most similar (due to sharing the same therapy group and therapist). The therapy groups are nested within the ten school-year strata that also may have some effects and lead to

correlation between outcomes recruited from the same stratum. This has implications for how statistical inferences are generated. Any analysis model will need to allow for these correlations and any resampling method that is aiming to generate realistic sampling distributions for parameters of interest will need to reflect the way in which participants are recruited to the trial. Chapter 3 provides details of suitable analysis models for therapy group effects and a resampling approach that mimics the trial data generating process.

2.1.4 Measurements

Participants' characteristics were measured at baseline (pre-randomisation). An interview collected information on child demographics (i.e. age, gender and development), parent demographics (i.e. education, ethnicity and mental health status) and measures of socio-economic status (i.e. family income, housing type, whether the child receives free school meals and family structure). Child word reading ability was assessed using the age-based standardised British Reading Ability Scale II - word reading. This is an individually administered test of the child's ability to read single words (Elliott et al., 1996b). Parent's mental health was measured using the General Health Questionnaire (GHQ) – 12 (Goldberg, 1972a). The total GHQ score range is 0-36 using a Likert scoring method.

Child antisocial behaviour was measured at baseline (time point 1) and one year after randomisation (time point 2) using the interview measure of parent account of child symptoms (PACS: (Taylor et al., 1986) as the primary outcome. This well validated semi-structured interview used investigator-based criteria to assess the frequency and severity of antisocial behaviours such as lying, tantrums, rudeness, disobedience, destructiveness and aggressiveness. The final score is yielded by taking the average of the frequency score and the severity score with a continuous range from 0 to 3. Child hyperactivity was also measured using the PACS interview. The latter outcome is not the focus of the investigation reported in this thesis. However, it is thought to be predictive of the outcome of interest (antisocial behaviour) and the measure is used when it comes to handling missing values in antisocial behaviour (for more details, see Chapter 3, Section 3.3).

Parenting behaviour was also measured at time points 1 and 2 via multi-informant and multi-method approaches such as semi-structured interviews (informed by investigator and parent), an expressed emotion interview (investigator rated), questionnaires (parent

completed) and direct observation (investigator rated). The semi-structured interview developed by Michael Rutter and colleagues (Rutter et al., 2011, Dowdney et al., 1984, Dowdney et al., 1985) was used to measure parent behaviours including *play, praise, rewards, consequence, timeout* and *harsh discipline,* which were scored from 0 to 4. *Expressed emotion* (EE) was a qualitative measure of the parent's emotion expressed towards the child and was measured by the investigator during the interview. The *parental warmth* score and the *parental criticism* score were derived from the EE interview using the Camberwell Family Interview criteria (Vaughn, 1989, Brown et al., 1972) with a four-point scale (0-1-2-3) reflecting the extent of expressed emotion. The Parent Practices Questionnaire (Webster-Stratton et al., 2008) was used for measuring *positive encouragement, harsh parenting* and *inconsistent parenting*. Parents were also directly observed following the procedure of the CPPRG (Conduct Problems Prevention Research Group, 1999). Fifteen-minute structured play tasks (free play, parent-directed task, and parent instructs the child to tidy away the toys) were given to the mother and child at home and videotaped. More details of measurement instruments and scales applied in the SPOKES study are provided in the SPOKES Data User Guide in the Appendix I.

2.2 Overview of CPT study

2.2.1 Participants

The study took place from 1995 to 1999 in four NHS child and adolescent mental health services (centres): Croydon, Brixton/Belgrave/Camberwell, St George's (all in south London) and Chichester (West Sussex). Eligible children were all those aged 3 to 8 years who were referred for antisocial behaviour to their local multidisciplinary child and adolescent mental health service. Exclusion criteria were clinically apparent major developmental delay, hyperkinetic syndrome, or any other condition requiring separate treatment. Parents had to be able to understand English, consent to the trial and attend at group times. In total, the study included parents of 110 children who were eligible. Compared with population norms, the referred children's mean scores were above the 97th percentile for conduct problems.

2.2.2 Study design

The CPT trial was a waiting list controlled study to see whether a behaviourally based group parenting programme, delivered in regular clinical practice, is an effective treatment for antisocial behaviour in children. Clusters of parents were allocated to the intervention or the

waiting list control arm by date of mental health service referral. Each cluster consisted of eligible referrals during a consecutive three-month period and all referrals were allocated to the same arm of the trial. Allocation was determined by date of receipt of the referral letter. The overall ratio of intervention to control clusters was 2:1. In total, seventy-three parents were allocated to the intervention arm and thirty-seven parents to the control arm. Both patients and researchers were blind to the allocation. Table 2-4 lists the counts of participants in each cluster for the waiting list control and the intervention groups.

Table 2-4 Number of participants in each cluster by treatment groups for CPT study

Clusters	No. Control	No. intervention	Cluster size
I	0	7	7
II	0	5	5
III	0	6	6
IV	0	7	7
V	2	0	2
VI	0	6	6
VII	0	7	7
VIII	0	8	8
IX	11	0	11
X	0	6	6
XI	0	6	6
XII	0	4	4
XIII	6	0	6
XIV	0	8	8
XV	0	3	3
XVI	6	0	6
XVII	4	0	4
XVIII	8	0	8
Total	37	73	110

2.2.3 Intervention

Parents allocated to the intervention group received the “Incredible Years” videotape parent training programme (Webster-Stratton and Hancock, 1998). Parents of six to eight children were seen as a group for two hours each week over thirteen to sixteen weeks. The children did not take part in the programme and no other treatment was given. The IY programme covered *play, praise and rewards, limit setting,* and *handling misbehaviour*. In each session, two group leaders showed videotaped scenes of parents and children together, which depict “right” and “wrong” ways of handling children. Parents discussed their own child's behaviour and were supported while they practised alternative ways of managing it. Each week, tasks were set for parents to practise at home and telephone calls made to encourage progress.

As shown in Table 2-5, therapy groups were nested within clusters in the intervention arm. This indicates that CPT data had a hierarchical structure: parents (level 1)-therapy groups (level 2)-clusters (level 3) in the intervention arm, and parents (level 1)-clusters (level 2) in the control arm. Again, this cluster randomisation design feature has implications for the choice of relevant analysis models and resampling procedures that aim to mimic the trial data generating process. I return to this point when justifying my choice of bootstrapping approach (Chapter 3 Section 3.2.4) and in particular when using the CPT data (Chapter 5).

Table 2-5 Counts of participants in each therapy group by clusters for CPT study

Clusters	Therapy Groups														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	*NA*	Total
IY intervention Group																
I				7												7
II												5				5
III					6											6
IV									4	3						7
VI						6										6
VII	7															7
VIII		3					5									8
X													6			6
XI											6					6
XII			4													4
XIV								8								8
XV														3		3
Subtotal	7	3	4	7	6	6	5	8	4	3	6	5	6	3	0	73
Waiting list control Group																
V															2	2
IX															11	11
XIII															6	6
XVI															6	6
XVII															4	4
XVIII															8	8
Subtotal															37	37
Total	7	3	4	7	6	6	5	8	4	3	6	5	6	3	37	110

NA: Not Applicable

2.2.4 Measurements

The measures in CPT were collected using multi-informant multi-methods. Measures were taken from parents on entry to the trial (time point 1) and after completion of the intervention or waiting list period (time point 2). The interval between the two time points was five to seven months. Child demographics (age, gender and development), parent demographics (education, ethnicity and mental health) and social economic status (family

income, housing, family structure and eligibility for free school meals) were measured at baseline. Unlike the SPOKES trial, CPT measured parents' mental health using the BECK Depression Inventory 21 (Beck et al., 1961), which includes twenty-one questions. The scoring method of each item was in ordinal order 0-1-2-3, and the total score ranged from 0 to 63.

In the same way as the SPOKES trial, child antisocial behaviour was measured at time points 1 and 2 using the parent account of child symptoms (PACS) interview as the primary outcome measure for antisocial behaviour in the CPT trial. Information about child hyperactivity was also collected in this interview.

Similar to the SPOKES trial, the CPT parenting behaviour was measured at time points 1 and 2 via multi-informant and multi-method approaches such as semi-structured interview (Dowdney et al., 1985), expressed emotion interview (Vaughn, 1989) and direct observation (Conduct Problems Prevention Research Group, 1999). These measurement methods measured positive parenting practice (*interview creative play, expressed warmth,* and *directly observed positive parenting*) and negative parenting practice (*interview smacking, expressed criticism,* and *directly observed negative parenting*). More details of measurement instruments and scales applied in the CPT study are provided in the CPT Data User Guide in the Appendix I.

2.3 Overview of HCA trial

2.3.1 Participants

The HCA trial has been conducted in two contrasting local authorities: a disadvantaged inner London Borough (Hackney) and a South West city (Plymouth). In these two locations, recruitment was conducted in two ways: first by a population (5-7 year olds) based screen in schools and secondly by seeking referrals from interested parents and teachers. Parents and/or teachers completed the SDQ or DSM questionnaire for assessing child antisocial behaviour. Participants were eligible to take part based on the following criteria: 1) Children met the screen cut-off: SDQ ≥ 3 or DSM ≥ 5; 2) Parent's ability to speak functional English; 3) Interest in taking part in the study; 4) Child score equal or above 0.7 on the Parent Account of Child Symptoms, Disruptive Behaviour scale; 5) Child free of global

developmental delay; and 6) Child score equal or above 70 on the British Picture Vocabulary Scale. At the end, 213 parents from the eligible families took part the trial.

2.3.2 Trial Design

The HCA trial was a randomised controlled trial with four trial arms (factorial design): 1) the "Incredible Years" (IY) only, 2) the literacy intervention only, 3) IY and literacy interventions combined, and 4) the control group. The trial took place between February 2008 and March 2012, assessing the effectiveness of three parenting programmes to reduce anti-social behaviour in primary school children living in a disadvantaged inner London Borough and a South West city.

Eight recruitment cohorts were formed based on the date when the participants enrolled into the trial. Within each recruitment cohort, participants were individually randomised to one of the trial arms. The overall ratio of the four treatment arms was 1:1:1:1. In total, among 213 participants, 56 were randomised to the IY only group, 53 to the literacy only group, 50 to the combined group, and 54 to the control group. Table 2-6 lists the numbers of participants in each trial arm by recruitment cohort.

Table 2-6 Number of participants in each trial arm by recruitment cohort for HCA trial

Recruitment Cohorts	No. IY	No. Literacy	No. Combined	No. Control	Total
I	5	0	5	1	11
II	6	12	22	14	54
III	3	8	13	2	26
IV	8	0	3	11	22
V	9	12	0	0	21
VI	9	10	7	16	42
VII	9	4	0	5	18
VIII	7	7	0	5	19
Total	56	53	50	54	213

However, different recruitment cohorts may have different ratios and/or different lists of allocation groups. There was a four-year overall plan laying out which interventions would be available; this was determined prior to any cases being randomised. The general rules were that recruitment cohorts should be determined within a month of the case being eligible for the trial, so that they should not wait for too long before knowing which treatment they were going to receive. This meant that the numbers varied according to how many joined in any given month. Some recruitment cohorts are considerably bigger due to

successful screening at those times. To handle this specific design feature and enable unbiased estimates of group effects, recruitment cohorts need to be conditioned on in the statistical analysis. More details of the HCA analysis model are provided in Chapter 5.

2.3.3 Interventions

The HCA trial provided three active interventions (IY, literacy, and combined) and one control. The IY intervention was the same intervention as in the CPT trial. The combined intervention was the same intervention as in the SPOKES trial. The literacy only intervention was the literacy part of the combined intervention. The control was service as usual/'signposting', which provided information to parents about services that were appropriate for concerns they raised about their child.

Table 2-7 Number of participants in each therapy group by recruitment cohorts and intervention arms for HCA trial

	Recruitment Cohorts								
Therapy Groups	I	II	III	IV	V	VI	VII	VIII	Total
A	5								5
D		2							2
F			3						3
H				5					5
J					3				3
K						4			4
L						5			5
M							2		2
N		4							4
T				3	6				9
Z							7		7
AC								7	7
IY Total	5	6	3	8	9	9	9	7	56
E	3								3
G			3						3
I					3				3
P	9								9
S			5						5
V					9				9
W						5			5
X						5			5
AA							4		4
AB								7	7
Literacy Total	12		8		12	10	4	7	53
B	5								5
C		4							4
O		18							18
Q			5						5
R			8						8
U				3					3
Y						7			7
COMBI Total	5	22	13	3		7			50
Control	1	14	2	11		16	5	5	54
Total	11	54	26	22	21	42	18	19	213

2.3.4 Measurements

The measures collected in the HCA trial consisted of a mixture of questionnaires, interviews and observations carried out in the child's home or in the school at baseline (pre-randomisation, time point 1) and 9 to 11 months after randomisation (time point 2). The measurement timing of HCA matches with SPOKES but the interval between time points 1 and 2 was shorter in CPT (5-7 months).

Similar to the SPOKES trial, the socio-demographic data were collected at baseline using a semi-structured interview which included details of the children's age, gender, development, eligibility for free school meals and family structure; parents' ethnicity and education. The national Statistics Socio-Economic Classification was used to assess parents' employment. In particular, the Depression Anxiety Stress Scale 21 (DASS-21) was used to measure mother's mental health in three dimensions (depression, anxiety and stress). The DASS-21 consists of three seven-item self-report scales that measure depression, anxiety and stress. A four-point severity scale (0-1-2-3) measures the extent to which each state has been experienced over the past week. The total DASS-21 score ranged from 0 to 21 for each dimension.

Child's antisocial behaviour was the primary outcome in this trial. As in the SPOKES trial and the CPT study, PACS was used to measure child antisocial behaviour. The measure was also used to assess the parent's detailed account of the severity and frequency of the child's restlessness and inattention (ADHD symptoms). The child's hyperactivity was not the primary outcome of this project and it was only used for predicting the missing values of the child's antisocial behaviour (see later in Chapter 5).

The HCA trial used semi-structured interviews (Dowdney et al., 1984), expressed emotion interviews (Vaughn, 1989), the Alabama Parenting Questionnaire (APQ) (Shelton et al., 1996) and direct observation (Conduct Problems Prevention Research Group, 1999) to measure parenting behaviours. The HCA trial measured a list of positive parenting practices including interview *creative play, praise* and *rewards, expressed parental warmth*, *Alabama positive parenting*, and directly observed *positive attends to child*. It also measured a set of negative parenting practices such as *smacking, expressed parental criticism*, *Alabama negative parenting,* and directly observed average negative parenting.

A general introduction to the measurement of child behaviours and parenting practises has been provided previously in Sections 1.2.2 and 1.2.3 of the thesis respectively. Further details of the measurement instruments and scales applied in each of the trials can be found in the measurement scale summary table of the Data User Guide in Appendix I.

2.4 Measurement harmonisation of three parenting studies

The three studies (SPOKES, CPT and HCA) of parenting programmes provide a rich data source for evaluating the mechanisms by which parenting interventions are thought to operate. These studies all aimed to improve child behaviour by targeting parenting practices. They investigated similar parenting programmes and employed compatible measures of child and parent outcomes. Therefore, pooling data from the three studies for an integrative data analysis should be feasible and is of interest here. I return to the methodological challenges posed by pooled data sets and possible solutions in Chapter 5. This section focuses on data preparation steps aimed at maximising the amount of data available for such future pooled analyses. In particular, baseline demographic and behaviour measures, measures of parenting behaviours and child outcomes require some harmonisation to enable pooling of data sets.

2.4.1 Steps involved in pooling multiple trial data sets

I conducted the following three steps to enable pooling of trial data sets:

Step1 – Identification of the underlying concepts measured in all trials:

The term *concept* here refers to an unobserved (latent) variable that holds the same meaning across trials despite there being differences in assessment instruments or scales. In the case of our parenting programme trials, a list of concepts measured across three trials is given by:

a. Child's demographics at baseline: child's gender, age;
b. Child's development at baseline: child's reading ability;
c. Parent's demographics at baseline: parent's ethnicity, education;
d. Parent's wellbeing at baseline: parent's mental health;
e. Family structure and socio-economic status at baseline: single parent, family size, child eligibility for free school meals;
f. Parent's parenting practices at baseline and time point 2: positive parenting practices, negative parenting practices;

g. Child's behaviours at baseline and time point 2: child's conduct disorder behaviours, child's hyperactivity;
h. Aspects of therapy: parent training session attendance, therapy groups;
i. Features of trial design: recruitment cohorts (of HCA) / school-year strata (of SPOKES) / clusters (of CPT).

Step 2 – Harmonisation of measures across trials:

A specific concept may be measured using the same instrument and scale with the same standardisation and coding methods across trials. In this ideal scenario, the measures can be combined directly without any harmonisation. In our parenting programmes trial data sets, the measures of child's gender, child's conduct disorder behaviour at baseline and time point 2 are examples of this scenario. Child's gender was obtained from a semi-structured interview as a binary variable. Child conduct disorder behaviour was measured using the PACS interview with the same items (interview questions) and rules used to construct conduct disorder behaviour scores across three parenting trials.

The same concept may be measured across trials but using different instruments. For example, parent's depression was assessed using three different questionnaires across three trials, i.e. the General Health Questionnaire 12 (GHQ-12) in SPOKES, the Beck Depression Inventory 21 (BDI-21) in CPT and the Depression Anxiety Stress Scales 21 (DASS-21) in HCA. To generate a commensurate measure, I standardised the depression score using the corresponding population means and standard deviations extracted from the published literature. The standardised score (z-score) was calculated using the formula: (raw score – population mean) / population standard deviation. The key of this harmonisation method is to find reliable literature sources that report summary statistics for the various measures in a reference population comprising those targeted in the trials. Influential factors on depression score, such as age, gender and region, were considered as the criteria to identify the corresponding population if possible. For SPOKES GHQ-12, the reference population was all females aged 16 – 65 in 1997, with mean and standard deviation of 11.50 and 5.08 respectively based on the Health Survey for England (Pevalin, 2000); for CPT Beck-21, the reference population norm was calculated from 4481 individuals from fourteen UK national samples with an age range from 15 to 70.1 and for both men and women. The UK national population BDI mean and standard deviation were 7.25 and 5.85 respectively (van Hemert et al., 2002); for HCA DASS-21, a sample of 1,794 members of the general adult UK population

(979 female, 815 male) was used to estimate the population mean and standard deviation, which were 2.83 and 3.87 respectively (Henry and Crawford, 2005). I refer to this method as *population z-score standardisation*.

There is also the situation that the same concept was measured using different scales across trials. For example, parent's education was measured as a binary variable with two categories (no education after 16 and further education after 16) in the SPOKES and CPT trials; while it was measured in three categories (educated to 16, educated to 18 or technical qualifications and educated to degree level) in the HCA trial. In this case, I combined the last two education categories together and created a binary parent's education variable for the HCA trial to ensure that a logically equivalent scale was used across trials. This harmonisation method is referred to in this study as *item re-categorisation*.

In addition, attention should be paid to the standardisation method (the form) of the measurements. In certain cases, it is hard to distinguish the raw score and the standardised score across different trials. Here, I use the child's reading ability measure as an example. Child's reading ability was assessed across three parenting trials. The raw score was recorded in SPOKES, but the age standardised score was derived and recorded in the CPT and HCA trials. Without thorough comparison of the distribution of the values across trials and careful checking of the variable label, the unit difference of the child's reading ability measures among trials is not easy to detect. To achieve *score conformance*, I calculated the age standardised British Reading Ability score using the BAS II manual based on the raw score and the age information collected in the SPOKES trial.

Different coding methods may be applied to the measures also. For instance, in the SPOKES and CPT trials, expressed parental criticism was coded as 0=no criticism; 1=very little criticism; 2=moderate criticism; 3=quite a lot of criticism; 4=a lot of criticism throughout, but a reverse coding was used in HCA, where 4=no criticism; 3=very little criticism; 2=moderate criticism; 1=quite a lot of criticism; 0=a lot of criticism throughout. Therefore, parental criticism of the HCA trial was recoded in order to eliminate the mismatching of the measures at the data pooling step. This process is called *coding reconcilement*.

In summary, this crucial harmonisation step requires thorough checking of the measures at the item level and careful inspection of the instrument, scale, standardisation and coding of the measurements. Application of the harmonisation approaches (population z-score standardization, item re-categorisation, score conformance, and coding reconcilement) leads to comparable measures of the concepts across the contributing trials.

Step 3 – Error checking:

After obtaining the harmonised measures, their distributions should be assessed to check for errors that might have occurred in step 2 of the data manipulation. I compared the distributions of the original data with that of the harmonised data to make sure all the differences were expected. Furthermore, comparison of the harmonised measures across contributing trials was conducted, as it helps to identify errors and might hint at the existence of heterogeneity across trials. Common techniques such as descriptive statistics and plots were employed here to compare the distributions of the measures.

Following the completion of the above checking procedure, data were pooled. The three trials' data were appended one after another and the pooled data included 415 observations, amongst which 112 were from SPOKES, 110 from CPT, and 213 from HCA. A trial variable was created to indicate which trial the data originally belonged to and a new ID variable was also created to distinguish individual observation among all contribution trials. In addition, therapy group (available in the treated arms only) and recruitment cohorts / school-year strata / clusters were coded in such a way that there were no duplicate values between trials. Then, identical measures (with or without harmonisation) were pooled together and formed one variable across three trials. After all these variables had been created, clear and informative labels were given to the variables and the category values respectively. The data pooling was conducted using SAS 9.3 (SASInstituteInc, 2011) software and the pooled data were also converted to STATA (StataCorp, 2011) format and SPSS (IBMCorp, 2011) format for future use.

2.4.2 Identifying the representative measure for the parenting practices

In this section, I will set up a selection strategy for the purpose of constructing a set of representative and consistent measures of parenting practices that will be investigated as

putative mediators in later mediation analyses using the SPOKES data (Chapters 3 and 4) and the pooled data from the three trials (Chapter 5).

As described in Chapter 1, Section 1.2.1, promoting positive parenting and reducing negative parenting are the two targets of the IY parenting intervention. However, positive and negative parenting practices are broad concepts, so that the measure of these two concepts covers a range of sub-concepts such as creative play, praise, rewards for positive parenting, and smacking and criticism for negative parenting. Since different measurement methods have been applied for measuring parenting practices (Sections 2.1.4, 2.2.4 and 2.3.4) in different trials, each trial provided a slightly different combination of sub-concepts measures. This inconsistency of measures has become an obstacle for performing mediation analyses using data from the three trials. There is, therefore, a definite need to select a shortlist of meaningful parenting practice sub-concepts that were measured in all three trials. To accomplish this purpose, I used the following strategy:

a. List all parenting practice sub-concepts that were measured using different measurement methods for each trial (see Table 2-8). In total, there were 80 variables to be considered in the three trials.
b. Identify the sub-concepts that were measured in all three trials (see **bold** and *italic* text in Table 2-8). This step greatly narrowed down the list of sub-concepts. Thus necessary checking activities were performed in step c.
c. Check the correlation between the selected sub-concept and the individual trial residual (non-selected) sub-concepts. Strong correlation suggests that the selected sub-concept is a good indicator of the non-selected sub-concept. For example, the interview smacking was selected as a representative measured of negative parenting practice because it was measured in all three trials. Although interview harsh discipline was not in the sub-concepts shortlist due to its absence in CPT, it was strongly associated with smacking measure in SPOKES and HCA with correlation coefficients of 0.78 and 0.85 respectively. This finding indicates that dropping harsh parenting will not result in the loss of much information.

Table 2-8 Parenting practices sub-concepts measured in each trial

Parenting Practices Concepts	**Measurement Methods**	**Parenting practices sub-concepts**		
		SPOKES	**CPT**	**HCA**
Positive Parenting	Interview	***1. creative play*** 2. praise 3. rewards	***1. creative play*** 2. NA 3. NA 4. parental coping	***1. creative play*** 2. praise 3. rewards
	Questionnaire	1. PBQ* positive encouragement	NA	1.Alabama Positive Parenting subscale 2.Alabama Involvement subscale
	Expressed Emotion	***1. expressed warmth*** 2. NA	***1. expressed warmth*** 2. NA	***1. expressed warmth*** 2. number of positive comments
	Observation	***1. positive attention*** 2. seek cooperation 3. facilitation 4. alpha commands	***1. positive attention*** 2. seek cooperation 3. facilitation 4. alpha commands	***1. positive attention*** 2. seek cooperation 3. facilitation 4. alpha commands
Negative Parenting	Interview	***1. smacking (tap or smack frequency)*** 2. harsh discipline	***1. smacking (tap or smack frequency)*** 2. NA	***1. smacking (tap or smack frequency)*** 2. hash discipline
	Questionnaire	PBQ* harsh parenting	NA	Alabama Corporal Punishment subscale
	Expressed Emotion	***1. expressed criticism*** 2. NA	***1. expressed criticism*** 2. NA	***1. expressed criticism*** 2. number of negative comments
	Observation	***1. negative affect*** 2. parental criticism 3. "don't" commands 4. impossible commands 5. beta commands	***1. negative affect*** 2. parental criticism 3. "don't" commands 4. impossible commands 5. beta commands	***1. negative affect*** 2. parental criticism 3. "don't" commands 4. impossible commands 5. beta commands
Limit Setting	Interview	1. consequences 2. time out 3. NA 4. NA	NA	1. consequences 2. NA 3. aversive discipline 4. calm discipline
	Questionnaire	1. PBQ* inconsistent 2. PBQ* appropriate discipline	NA	1. Alabama Inconsistent Discipline 2. NA 3. Alabama Poor Supervision subscale
	Observation	***Limit setting (seek cooperation/negative commands)***	***Limit setting (seek cooperation/negative commands)***	***Limit setting (seek cooperation/negative commands)***

*PBQ: parenting behaviour questionnaire.

d. For measures obtained via direct observation, there is no standard approach available for creating positive parenting and negative parenting scales. An informal factor analysis was conducted to identify the influential factors of parenting practices. This gave three factors. The first, with 29% of the variance, was Alpha

commands, don't, impossible commands, and negative affect; the second, with 19% of the variance, was facilitate, positive attention, and beta commands; the third was seeking cooperation. On theoretical grounds, it would seem likely that positive parenting should be represented by positive attention, negative parenting by negative affect, and limit setting by some proportion between seeking cooperation and negative commands. More details about the exploration of direct observed data can be found in Appendix I.

e. Generally speaking, appropriate limit setting (such as consequences, calm and appropriate discipline) and inappropriate limit setting (such as inconsistent and aversive discipline) can be categorised broadly as positive and negative parenting respectively. Table 2-8 listed limit setting parenting practices in a separate category because they were investigated separately in previous studies of the PTMO and Triple P parenting interventions (see Chapter 1, Table 1-2). However, in this project, only the limit setting score calculated from the direct observation measurement was available in all three trials. It is also known that direct observation methods might not be reliable because the observations in a short period of time (one to two hours) may not be typical of what is going on at home. For these reasons, I did not consider limit setting practices as the putative mediators of interest. Nevertheless, they were included in predicting the missing values of the intermediate and outcome variables (see Chapter 3, Section 3.3.1).

Application of the above strategy resulted in a list of sub-concepts for all three trials under each measurement method (see Table 2-9). These sub-concepts are the set of putative parenting mediators that we are going to examine in the following chapters.

Table 2-9 Pooled measurements of parenting practices

Concepts	Measurements	Sub-concepts
Positive Parenting	Michael Rutter Interview	Creative Play
	Expressed Emotion Interview	Warmth
	Direct Observation	Observed Positivity
Negative Parenting	Michael Rutter Interview	Smacking
	Expressed Emotion Interview	Criticism
	Direct Observation	Observed Negativity

2.5 Discussion

2.5.1 Differences among the three parenting trials

Although the parenting trials included in the project are very similar and comparable (see description at the beginning of Section 2.4), the differences in outcome distributions between trials indicate the existence of trial effects. Firstly, the antisocial behaviour severity of the study population is different: SPOKES children's antisocial behaviours were above one standard deviation over the population mean. Compared to SPOKES, the CPT study population was more severe. CPT children's antisocial behaviour mean score was above the 97th percentile of the population norms. The HCA trial had the mildest study population among the three trials. HCA children's antisocial behaviour scores were just above the population mean. Secondly, the child outcomes were collected at different times in the three trials. SPOKES time point 2 was one year after randomisation. Similar to SPOKES, HCA time point 2 was nine to eleven months after randomisation, which was four months after the completion of the intervention. However, CPT time point 2 was five to seven months after enrolment. Consequently, the effect of trials should be taken into account in the pooled analysis. In addition, different trial populations may respond to the same influential factor (explanatory variable) differently, so that potential effect moderation across trials should also be allowed in the pooled analysis. Details of the statistical modelling of the pooled data analysis are described in Chapter 5.

2.5.2 Comparison with complex modelling technique of data-pooling

Hussong, Curran and Bauer (Hussong et al., 2013) introduced a complex modelling approach to create a commensurate scale using item level data from multi-item measures of multiple studies for integrative data analysis (pooled data analysis). The approach included four key steps: Step 1 involved preliminary feasibility checking. The approach requires that common items be present in at least pairs of studies, so that sufficient pairs can link the measurement across studies. Step 2 involved selecting an item set. Based on the theoretical group, a pool of candidate items were initially identified as relevant elements of the latent factor. After performing exploratory factor analysis using this item pool, a final uni-dimensional set of items was selected for the factor of interest. Step 3 involved developing a measurement model. The step 2 factor model with the selected items and the predictor of the factor mean and variance was fitted to work out conditional distribution of latent factor given individuals' observed data. Step 4 was scoring. The factor scores for the individual participants were

predicted by the expectation of the conditional distribution of the latent factor (Empirical Bayes prediction).

However, this commensurate scale generating approach has several limitations compared with the harmonisation approaches proposed in Section 2.4.1. Firstly, this commensurate approach is only applicable if items have been observed in at least pairs of studies. This is not always the case in this project. For example, parental depression was measured using unique instruments across all three trials (GHQ-12 in SPOKES, BDI-21 in CPT and DASS-21 in HCA). Secondly, adopting this commensurate approach in the context of mediation analysis excessively increases modelling complexity. Specifically, in order to avoid the pitfalls of two-stage approaches (Skrondal and Laake, 2001) – that is, generating the commensurate scale in the first stage and fitting the mediation analysis model in the second stage – complex SEMs need to be fitted for combining both the scoring model and the analysis model in a single stage. This area is still under development and is not the interest of the current project. Finally, the commensurate approach does not exploit external population's data. In contrast, the proposed harmonisation approach in the parental depression example borrowed information from the reference population. In sum, the measurement harmonisation approaches proposed in this project aim at providing simple, straightforward and meaningful pooled measures across contribution trials. The proposed approaches can cover further scenarios compared to the complex commensurate approach.

Chapter 3 Mediation Analysis under Observed Confounding and in the Presence of Missing Values

3.1 Introduction

As discussed in Chapter 1, the traditional linear regression approach proposed by Baron and Kenny (Baron and Kenny, 1986) has been applied to investigate mediation of the effects of parenting programmes on child behaviour problems by improved parenting practices. However, as outlined in Chapter 1, such a regression approach might be subject to bias due to the confounding of the relationship between the mediator and the outcome, missing values in the mediator, outcome and covariates, existing hierarchical data structures and inappropriate distributional assumptions. In this chapter, I will propose a new mediation analysis method that is valid under observed confounding, in the presence of missing values and allowing for hierarchical data structures. The new methodology will be illustrated with the motivating example of the SPOKES RCT. I will start by further discussing these statistical challenges in the context of IY parenting programme trials in the following sections.

3.1.1 Measured confounders between parenting mediator and child outcome

It has been mentioned in Section 1.4.4 of Chapter 1 that confounding between the mediator and outcome might be present in RCTs and that the existence of such confounding leads to biased mediation effect estimates. In trials of IY parenting programmes, it is true that confounding of the parenting practice mediator and child behaviour outcome relationship cannot be ruled out. In addition, the literature on the underlying psychological theory and experimental research suggest that observed baseline variables such as child gender, ethnicity and intelligence, family poverty, family structure, parental mental health and parental education affect both parenting practice and child antisocial behaviour (Bloomquist and Schnell, 2002, West et al., 2000, Webster-Stratton and Hammond, 1997). Consequently, the existence of such common causes of parenting practice and child antisocial behaviour leads to a biased estimate of the effect of parenting practice on child outcome if the estimator is based on these two variables alone.

3.1.2 Missing data in trials of IY parenting programme

In trials of parenting programmes, missing values might be present in baseline confounders, intermediate parenting outcomes and/or distal child antisocial behaviour outcomes. Taking

the SPOKES trial as an example, I found that more than 40% of the observations have at least one missing value in a list of baseline confounding variables (see Table 3-1); 9% of the observations have a missing value in the child antisocial behaviour outcome; up to 26% of the observations have missing values in the putative parenting practice mediators. Although the percentage of missing values might not be high in each variable, the complete cases based on availability of measures for a list of confounders, a mediator and the child outcome are less than 50% of the total cases (the total rows / observations of the dataset).

Table 3-1 Number of observations in the SPOKES trial with different numbers of missing values for a list of seven putative baseline confounding variables

No. of Missing Values	No. of observations	Percent	Cumulative %
0	64	57.1	57.1
1	34	30.4	87.5
2	11	9.8	97.3
3	3	2.7	100
Total	112	100	

Note: The list of putative baseline confounders includes child ethnicity, mother's education and depression, family income, eligibility for free school meals, lone parent and family size

As discussed in Chapter 1, the CC mediation analysis is inefficient. In the SPOKES trial, this approach only includes half of the cases and the rich information in the non-completed cases is discarded completely. The assumption that missing values only depend on the mediator, outcome, treatment received and the measured confounders included in the analysis model is implausible, and thus may lead to biased estimate of the causal effects of interest.

3.1.3 Hierarchical structure of IY parenting programme data

The IY parenting programme is a group training intervention. That is, parents who receive the IY parenting intervention will be trained in a group and discuss their parent-child experiences within the same group. In the SPOKES trial introduced in Chapter 2, randomisation was conducted within school-year strata and in each school-year stratum parents randomised to the intervention arm received IY parenting intervention in groups. Thus, the SPOKES trial has a hierarchical data structure with three levels: Level 1 - individual participants (parents); Level 2 - parents allocated to IY are nested within therapy groups, and

Level 3 – either individual control parents or therapy groups in the IY arm are nested within school-year strata. Similarly, in the CPT trials, which used a cluster randomised design, trial participants (level 1) allocated to IY were nested within therapy groups (level 2), and therapy groups were nested within randomisation clusters (level 3). The HCA trial also has a three-level hierarchical data structure: participants (level 1) – therapy groups (level 2) – recruitment cohorts (level 3). As introduced in Chapter 1, Section 1.4.4, Baron and Kenny's approach to mediation analysis might lead to biased standard error due to ignoring the ICC caused by the hierarchical structure. Therefore, development is required to account for specific trial designs in mediation analysis.

3.1.4 Non-normality in mediation analysis of IY parenting programme trials

Typically statistical inference consists of a point estimate of the parameter of interest accompanied by a confidence interval (CI) reflecting the precision of the estimate. From a frequentist point of view, a 95% confidence interval indicates that if a population were sampled repeatedly and a confidence interval calculated from each sample (the CI would differ from sample to sample), 95% of these confidence intervals would contain the population parameter of interest. When inferences are derived from parametric models, distributional assumptions might inform inferences at two stages: 1) The models' error terms (e.g. the noise component of a regression model) are assumed to follow a known distribution, typically a normal distribution with expectation zero, and/or 2) The sampling distribution of parameters is assumed to be normal, at least asymptotically (for large samples).

Sampling distribution of the indirect effect estimator: In the context of traditional linear regression mediation analysis, the most commonly used method to test an indirect effect is to divide the estimate of the indirect effect by its standard error and compare the resulting test statistic with a critical value from the standard normal distribution (Sobel, 1987). Confidence limits (upper and lower limits of the confidence interval) for the indirect effect are also typically based on critical values from the standard normal distribution. However, as discussed in Section 1.4.3 of Chapter 1, the causal indirect effect is estimated by the product term of two parameters ($\hat{\alpha}\hat{\beta}$). Even when the normality assumption of the sampling distribution of both parameters holds, the sampling distribution of the product term of two parameters can be skewed and has different values of kurtosis (Craig, 1936). Simulation

study results also suggest that confidence limits based on standard normal distribution critical values are imbalanced (as the distribution of the indirect effect is normal only in special cases) and the significant test of the indirect effect has statistical power and type I error rates that are too low (MacKinnon et al., 2004). Thus it would be preferable to generate inferences for indirect effects without relying on a normal assumption.

Regression error distribution: Traditional linear regression mediation analysis approaches rely on the assumption of normally distributed errors of both mediator model and outcome model to construct tests and confidence intervals for regression coefficients. In fact, suitably-defined parenting practice mediators (see Chapter 2) are discrete, while approximately scored on a continuous scale, and do not follow normal distribution. For example, the putative mediator *Expressed Warmth* was rated on a four-point scale: 0, 1, 2, and 3 to indicate the extent of warmth expressed by the parents from "no warmth" to "a great deal of warmth" in an increasing trend. For finite samples, confidence intervals for regression coefficients α, β, γ rely on the normality assumption for the error terms. Thus, while it might be reasonable to assume a linear relationship between such a mediator and the child outcome variable, it would be preferable to generate inferences that do not rely on normality assumptions about the error terms.

This chapter is split into major two parts: (I) Methodology development and (II) Application of the new methodology to analyse mediation in the SPOKES trial. Section 3.2 begins by reviewing the statistical methodology necessary for conceptualising and addressing the statistical challenges posed by traditional mediation analysis. I will later utilise aspects of existing analysis methods such as multiple imputation, linear mixed modelling/multilevel modelling and bootstrapping; thus I provide a brief introduction to these methods before proceeding to use them. Section 3.3 then proposes a new method for mediation analysis, referred to as the *MI-BT approach,* which is valid under less restrictive assumptions and can exploit all the information provided in our trials. Following the methodology development, part (II) of this chapter uses data from the SPOKES trial to illustrate the application of the new mediation analysis approach. Section 3.4 demonstrates the steps involved in setting up an MI-BT analysis for SPOKES and reports the findings. Section 3.5 concludes the chapter with a discussion of the advantages and disadvantages of the new mediation analysis approach.

3.2 Review of related statistical methodology

3.2.1 Handling measured confounders by conditioning

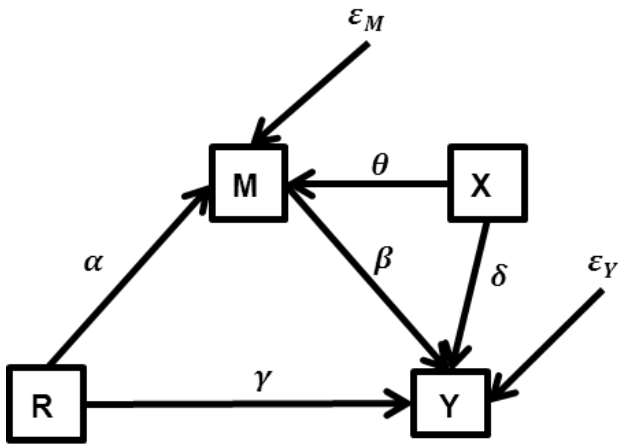
Bias due to measured confounders (see Section 3.1.1) can be avoided by including such confounders as explanatory variables in the regression models for the mediator and the outcome respectively and calculating OLS estimators. More specifically, the causal direct and indirect effects conditional on the values of measured confounders can be estimated by fitting the two linear regression models:

$$Y = i_1 + \mathbf{X}\boldsymbol{\delta} + \gamma R + \beta M + \varepsilon_Y \qquad \text{Equation 3-1}$$

$$M = i_2 + \mathbf{X}\boldsymbol{\theta} + \alpha R + \varepsilon_M \qquad \text{Equation 3-2}$$

where $Y, M, R, i_1, i_2, \varepsilon_Y, \ and \ \varepsilon_M$ hold the same meaning as in Equation 1-9 and Equation 1-10 of Chapter 1. $\mathbf{X}$ represents a set of r confounding variables (covariates), so that $\mathbf{X} = (X_1, X_2, \dots, X_r)$. The vectors $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ present a set of coefficients indicating the strength of the relationship between $\mathbf{X}$ and Y, $\mathbf{X}$ and M respectively. However, the values of $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ are not our research interest in this project. The coefficients $\alpha, \beta, and \ \gamma$ represent the strength of the relationships between the corresponding $Y, M, and \ R$ conditional on the observed confounders in $\mathbf{X}$. An RCT single-mediator diagram including the measured confounders is shown in Figure 3-1. Compared to the single-mediator model Figure 1-2 in Chapter 1, the major difference is that the effects of the measured confounders of the mediator – outcome relationship are now also modelled and thus $\alpha, \beta, \text{and } \gamma$ represent *conditional* mediation effects.

Figure 3-1 RCT single-mediation mediation model including measured confounders



The conditional causal effects of interest can be estimated by respective OLS estimators $\hat{\alpha}, \hat{\beta}, \text{and } \hat{\gamma}$ based on extended regression models, shown as Equation 3-1 and Equation 3-2. The conditional direct effect (DE) is estimated by $\hat{\gamma}$, the conditional indirect effect (IE) is estimated by the product of $\hat{\alpha}\hat{\beta}$, and the conditional total effect (TE) is estimated by $\hat{\alpha}\hat{\beta} + \hat{\gamma}$. The approach provides valid inferences under the assumptions that 1) the linear models hold (including linearity of the relationships between the mediator and the clinical outcome and the covariates and the outcomes), 2) there is no unmeasured confounding of the M- Y relationship and 3) there is no $R \times M$ interaction effect on Y.

3.2.2 Multiple imputation to deal with missing data

3.2.2.1 Classification of missing value generating mechanisms

Prior to elaborating on the Multiple Imputation (MI) approach for dealing with missing values, I will introduce Rubin's classification of missing value generating mechanisms, which helps to understand the missing data assumptions required by various analysis approaches. Rubin's classification of missing value generating mechanisms (Rubin, 1976, Little and Rubin, 2002) includes three types of missing mechanism, namely *Missing Completely at Random* (MCAR), *Missing at Random* (MAR) and *Missing Not at Random* (MNAR). These mechanisms describe relationships between measured variables and the probability of missing data. While these terms have a precise probabilistic and mathematical meaning, they are essentially three different explanations for why the data are missing. From a practical perspective, the mechanisms are assumptions required by different missing data handling techniques.

Missing Completely at Random (MCAR) specifies that missingness is completely unsystematic and that the observed data can be thought of as a random subsample of the hypothetically complete data. MCAR is a restrictive assumption, as it stipulates that the probability of a particular missing value pattern is not predicted by any observed or unobserved variables. Let us assume that a data set $\boldsymbol{D}$ consists of two parts: $\boldsymbol{D_o}$ is the observed part, $\boldsymbol{D_m}$ is the unknown (missing) part and $\boldsymbol{\Omega}$ is a binary matrix known as the missing data indicator matrix. The MCAR condition can be expressed by the relation $p(\boldsymbol{\Omega}|\boldsymbol{D_o}, \boldsymbol{D_m}) = p(\boldsymbol{\Omega})$.

Missing at Random (MAR) is often confusing and misleading because of the use of the word "random." In fact, it describes systematic missingness where the probability of a particular set of values being missing for an individual depends only on a set of observed variables but not the unobserved values themselves. The MAR condition can be expressed by the relation $p(\boldsymbol{\Omega}|\boldsymbol{D_o}, \boldsymbol{D_m}) = p(\boldsymbol{\Omega}|\boldsymbol{D_o})$. In this project, MI is implemented under MAR.

Missing Not at Random (MNAR) is the most problematic mechanism and can cause substantial bias. The MNAR mechanism describes that the probability of missingness is systematically related to the hypothetical values that are missing. It requires specialized analysis procedures (e.g., selection model, pattern mixture models) (Little, 1993) and can also occur when the "cause" of missingness is a measured variable that is omitted from the analysis model. The MNAR scenario will not be considered in this project.

3.2.2.2 A short overview of Multiple Imputation

In recent decades, Multiple Imputation (Rubin, 2004) has emerged as a convenient and flexible paradigm for analysing data with missing values. The basic idea of imputation is "filling in" missing data with plausible values (Schafer, 1999). In single imputation, only one estimated value is used to fill the missing value. In multiple imputation, each missing value is replaced by a list of $H > 1$ simulated values that produce H plausible versions of alternative complete data. Each of the complete data is analysed using standard methods and the results are later combined to produce estimates and confidence intervals that incorporate missing-data uncertainty.

A key feature of MI is that the imputation phase is operationally distinct from subsequent analyses. This feature allows differences between the imputation model and the analysis model. The imputation model does not intend to provide a parsimonious description of the data, nor does it represent structural or causal relationships among variables. The model is merely a device to preserve important features of the joint distribution (means, variances, and correlations) in the imputed values. The behaviour of repeated imputation inference when the imputation and analysis models differ has been investigated by Fay, Meng and Rubin (Fay, 1992, Meng, 1994, Rubin, 1996). Basically, when the imputation model is more general (i.e. makes fewer assumptions) than the analysis model, then MI leads to valid inferences with perhaps some loss of power. On the other hand, when the imputation model

makes more assumptions than the analysis model and the extra assumptions are plausible, MI intervals tend to be narrower than intervals derived purely from the analysis model, and they also tend to be conservative, with higher-than-nominal coverage probability. However, if these extra assumptions made in the imputation model are unwarranted, then the MI estimates will be biased. In practice, this means that an imputation model should reasonably preserve all the distributional features (e.g. associations) that will be the subject of future analyses.

The analysis results of the multiple imputed data were combined to produce a single MI estimator that is the simple average of the estimators calculated from H imputed data sets. This MI point estimator is unbiased under the MAR assumption (specifically all variables driving missingness have to be included in the imputation model). Generally, standard errors and confidence intervals for the combined estimator are constructed using Rubin's rules (Rubin, 2004), which take into account both the within- and the between-imputation variations. However, Rubin's rules are not applied in this project due to lack of theoretical support for combing MI with bootstrapping for generating statistical inference. More specifically, in this project, a non-parametric bootstrapping method is used to generate statistical inferences (see Section 3.2.4) in order to relax the distributional assumptions of the analysis but Rubin's rules were not developed in accordance with the underlying theory of bootstrapping in respect of estimating the variance of the estimator. Thus I only use the MI point estimator (the average of the estimators calculated from H imputed data sets) from the MI procedure, and the statistical inference (e.g. confidence interval of the estimator) is estimated using a non-parametric bootstrapping method.

The MI procedure used in this project is *Multiple Imputation by Chained Equation* (MICE) (Van Buuren and Oudshoorn, 2000). This is a flexible procedure which generates missing values based on a set of sequential multiple regression imputation models, one for each variable with missing values. A number of attractive features of MICE make it a flexible and practical tool for handling missing values. Firstly, MICE can handle different types of variables with missing values, including continuous, binary, ordered/unordered categorical variables. Secondly, a specific technique of MICE, i.e. *Predictive Mean Matching* (PMM), can be used for the imputation of skewed continuous, count and discrete variables (Little, 1988). Thirdly, MICE can also incorporate variables that are functions of other variables, such as

interaction terms. And finally, each variable can be modelled separately in the imputation step using a list of sensible and sound missingness predictors tailored specifically for this variable (White et al., 2011).

3.2.2.3 Comparison between MI and other techniques for dealing with missing values

MI analysis is commonly advocated for handling missing data in preference to CC analysis for several reasons. Firstly, MI analysis uses information from the incomplete cases, so that it is more efficient than CC analysis (Little, 1992). Specifically, in the context of mediation analysis, MI is able to use all available information in the mediators, the outcome and the measured confounders, whereas CC only analyses a subsample with no missing values in all the variables included in the analysis model. In particular, MI analysis gains efficiency when more variables are included in the analysis model. The more covariates (with missing values) are added into the analysis models, the more inefficient the CC analysis becomes. Secondly, it is believed that the MAR assumption made by MI is more plausible and the MI analysis is less likely to be biased, as more variables are included in the imputation model (Rubin, 1996). MI analysis is valid under the assumption of MAR, whereas CC analysis requires MCAR or a form of MAR where the variables driving missingness are included as covariates in the analysis model. In many scenarios, this assumption is considered as restrictive and less plausible. Additionally, carrying out CC mediation analysis explicitly makes one susceptible to harmful mistakes. For example, running CC regression analysis for the mediator model and the outcome model separately may lead to the use of different cases for the two models and thus to a biased estimate.

Another approach for dealing with missing data is *Inverse Probability Weighting* (IPW: (Seaman and White, 2013). This approach consists of a CC analysis with complete cases weighted by the inverse of their probability of having observed data (i.e. not being missing). Those who had a small chance of being observed are given an increased weight to compensate for those similar subjects who are missing. IPW separates the analysis model from the missing value generating model and assumes that a case's probability of being fully observed can be modelled as a function of observed variables (MAR). This might lead to bias reduction from the CC analysis. On the other hand, IPW only uses complete cases, whereas MI can use information from individuals with partially missing data: therefore, MI is more powerful than IPW.

Similar to MI, *Full Information Maximum Likelihood* (FIML) is another modern missing data procedure that is based on a sound theory and can produce efficient estimates and accurate measures of statistical uncertainty under MAR. Basically, the FIML estimation maximizes the sum of the log-likelihood functions for individual observations, including both complete and incomplete observations. Commonly, MI can be viewed as a computationally easier way to approximate the full-likelihood solution. When the imputation and analysis models are the same, full likelihood and MI with a very large number of imputations yield the same estimates (Collins et al., 2001). Auxiliary variables can be included in both MI and FIML procedures to improve efficiency and reduce bias provided the auxiliary variables are potentially correlates of missingness. When the imputation model uses auxiliary variables, MI is generally easier to implement than full likelihood.

To sum up, MI is a more efficient, less biased and relatively easier way to handle missing data under the assumption of MAR compared with other missing data procedures. MI can make use of all the available information in the parenting mediators, child outcomes and measured confounders for mediation analysis. Additionally, trials of parenting intervention used in this project measured various aspects of the parenting practices and child behaviour outcomes using multi-measurement methods such as questionnaires, interviews and direct observations. These additional measures provide a rich set of auxiliary variables that can be included in the imputation model to improve efficiency and reduce bias.

3.2.3 Multi-level modelling to account for hierarchical data structure

As pointed out in Section 3.1.3, trials of parenting intervention might have a hierarchical data structure. In the following paragraphs, I will describe the multilevel modelling approaches for trial data with hierarchical structure.

The linear mixed model is a popular statistical method for analysing hierarchical data (McCulloch, 2008). Mixed models are characterized as containing both fixed effects and random effects. The fixed effects are analogous to standard regression coefficients and represented by a set of parameters that explain changes in expected outcomes. In contrast random effects are random variables that follow distributions and are summarized according to their estimated variances and covariances. Random effects may take the form of either random intercepts or random slopes (random coefficients), and the grouping structure of

the data may consist of multiple levels of nested groups. As such, mixed models are also known in the literature as multilevel models or hierarchical linear models.

In the previous section, the regression coefficients in Equation 3-1 and Equation 3-2 are considered as fixed effects in the analysis models. I will now extend these regression models to include random effects in addition to the overall residual error to account for the hierarchical data structure.

Firstly, I will provide a brief review of the linear mixed model formula and parameter estimation. Linear mixed model in matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\delta} + \boldsymbol{\Psi}\mathbf{u} + \boldsymbol{\varepsilon} \qquad \text{Equation 3-3}$$

where $\mathbf{Y}$ is $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times p$ design/covariate matrix for the fixed effects $\boldsymbol{\delta}$, and $\boldsymbol{\Psi}$ is the $n \times q$ design/covariate matrix for the random effects $\mathbf{u}$. The $n \times 1$ vector of errors, $\boldsymbol{\varepsilon}$, is assumed to be multivariate normal with mean zero and variance matrix $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\mathbf{I}_{\boldsymbol{n}}$. The fixed portion of Equation 3-3, $\mathbf{X\delta}$, is analogous to the linear predictor from a standard OLS regression model with $\boldsymbol{\delta}$ being the regression coefficients to be estimated. For the random portion of Equation 3-3, $\boldsymbol{\Psi}\mathbf{u} + \boldsymbol{\varepsilon}$, we assume that $\mathbf{u}$ has *variance-covariance matrix* $\mathbf{G}$ and that $\mathbf{u}$ is orthogonal to $\boldsymbol{\varepsilon}$ so that

$$\mathrm{Var}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\mathbf{I}_{\mathbf{n}}\end{bmatrix} \qquad \text{Equation 3-4}$$

The random effects $\mathbf{u}$ are characterized by the elements of $\mathbf{G}$, known as *variance components*, that are estimated along with the overall residual $\boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$.

One of the most popular estimation methods is the maximum likelihood (ML) method, which is the default method of the STATA command *xtmixed*. The ML estimates are based on the usual application of likelihood theory, given the distributional assumptions of the model. Considering the combined error term $\boldsymbol{\Psi}\mathbf{u} + \boldsymbol{\varepsilon}$, we see that $\mathbf{Y}$ is multivariate normal with mean $\mathbf{X\delta}$ and $n \times n$ variance-covariance matrix $\mathbf{V}$, where

$$\mathbf{V} = \mathbf{\Psi G \Psi}' + \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}\mathbf{I_n} \qquad \text{Equation 3-5}$$

Defining $\boldsymbol{\eta}$ as the vector of unique elements of $\mathbf{G}$ results in the log likelihood

$$\ell(\boldsymbol{\delta}, \boldsymbol{\eta}, \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}|\mathbf{Y}) = -\frac{1}{2}\{nlog(2\pi) + log|\mathbf{V}| + (\mathbf{Y} - \mathbf{X\delta})'\boldsymbol{V}^{-1}(\mathbf{Y} - \mathbf{X\delta})\} \qquad \text{Equation 3-6}$$

which is maximized as a function of $\boldsymbol{\delta}, \boldsymbol{\eta}, and\ \boldsymbol{\sigma}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$.

Chapter 2 of Pinheiro and Bates' (Pinheiro and Bates, 2000) book provides details of the calculation of the ML estimators $\widehat{\boldsymbol{\delta}}$, $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$, $\widehat{\boldsymbol{\eta}}$ according to likelihoods in Equation 3-6. From the formula of the estimates, we know that $\widehat{\boldsymbol{\delta}}$ and $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$ depend on the estimate of the variance parameters in $\widehat{\boldsymbol{\eta}}$, so that different random component variance-covariance estimates give different fixed effect estimates. It has been proved that $\widehat{\boldsymbol{\delta}}_{\mathbf{ML}}$ is an asymptotically unbiased estimator of the fixed effect parameter (Robinson, 1991). The ML estimates $\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\varepsilon}}^{\mathbf{2}}$ is biased, as its divisor is n rather than $(n - p)$. This bias may be severe if the sample size is small compared with the number of fixed effect parameters. However, in my mediation analysis models, the focus is on estimating fixed causal mediation parameters. The random effects serve to account for the hierarchical data structure implied by the trial design but their variance parameters are not of interest in themselves. Thus I only consider the asymptotically unbiased fixed effects estimator $\widehat{\boldsymbol{\delta}}_{\mathbf{ML}}$ in my mediation approach.

3.2.4 Nonparametric bootstrapping for addressing non-normality

As reviewed in Section 3.1.4, the sampling distribution of the indirect effect is not necessarily normal. In addition, I did not wish to make a distributional assumption about putative mediators that are measured in a discrete scale. Therefore, a non-parametric bootstrap approach was pursued to generate confidence intervals and significance tests without making distributional assumptions regarding the errors in the regression models or the sampling distributions of estimators. In the following sections, I will briefly introduce the bootstrapping method and discuss how to use a nonparametric bootstrap approach to generate statistical inference for causal mediation effects of interest.

3.2.4.1 A brief review of bootstrapping

Bootstrapping is a well-established resampling method for assigning measures of accuracy and precision (e.g. bias, standard error and confidence interval) to statistical estimates (Efron and Tibshirani, 1993). The basic idea of bootstrapping is that the inferences about population parameters from the sample data can be generated by mimicking the sampling process and resampling the sample. The approach mimics what would happen if the sample were in fact the population and the bootstrap re-samples were repeated samples drawn from this population.

In its simplest form, bootstrapping generates a number of re-samples of the observed sample units by random sampling with replacement. The original sample is assumed to be a random sample from a target population. Under this re-sampling scheme, some of the original observations will appear once, some more than once, and some not at all. From each of the bootstrap samples (the re-samples) an estimator of population parameter is calculated and the resulting set of estimator values provides an empirical approximation of the sampling distribution of the estimator. Then, the properties of the estimator – such as its bias, standard error or a 95% confidence interval – can be approximated using the empirically generated sampling distribution. In this bootstrap procedure, no parametric model was specified. The only assumption is that the original sample data are identical and independently distributed according to an unknown distribution function, so that this bootstrap approach represents a non-parametric approach. The mathematical details of calculating the bootstrap confidence interval, standard error and bias for an estimator will be introduced in Section 3.2.4.3.

3.2.4.2 Bootstrapping hierarchically structured data

For data with a hierarchical structure, observations on level-1 units that belong to the same higher level unit (e.g. patients in the same therapy group) are correlated. Simply re-sampling level-1 units does not maintain the correlation structure and thus such re-samples would not mimic the data generating process. A more complex resampling strategy is required here to preserve the hierarchical structure. Davison and Hinkley (Davison and Hinkley, 1997) compared two bootstrap strategies for nested two-level hierarchical data structures in their book . Strategy 1 involves randomly sampling higher level groups with replacement followed by randomly sampling lower level data without replacement within groups (i.e. including all

the lower level units of that group if the size of the sample at lower level is equal to the number of sampling units in that group). Strategy 2 involves randomly sampling higher level groups with replacement followed by randomly sampling lower level data with replacement within groups. Their discussion centred on the first two moments (mean and variance) of the resulting estimators, suggested that strategy 1 is the preferred approach. Additionally, more recent simulation findings for the nonparametric bootstrap approach for three-level hierarchical data structures (Ren et al., 2010) agreed with strategy 1 and recommended a sampling approach of sampling with replacement at the highest level and including all the lower level units. Resampling should work well if the number of highest level groups is greater than or equal to 10, just as resampling homogeneous data works well if the number of samples is moderately large.

3.2.4.3 Calculating the bootstrap bias-corrected (BC) confidence interval, standard error and bias for an estimator

Suppose the original data are all complete cases, the bootstrapping approach discussed above provides 1000 bootstrap samples, in each of which I can run the proposed mediation analysis and get 1000 estimates of each parameter of interest. The 1000 estimates of each parameter form its empirical distribution. In the following paragraphs, I will summarise guidance from the literature on generating bootstrap inferences for mediation effects from this empirical distribution.

The bias-corrected confidence interval (CI_{BC}) was recommended for mediation effects, as simulation results show that the bias-corrected confidence interval has accurate Type I error rate and the largest power among many different confidence intervals evaluated (MacKinnon et al., 2004). The simple *percentile confidence interval* uses the percentiles of the bootstrap empirical distribution. For example, if the parameter of interest is θ, the estimator $\hat{\theta}^b$ is calculated from each bootstrap sample ($b = 1, \ldots, B$), where B is the total number of bootstrap samples. The 95% CI uses the percentile $\hat{\theta}^b(\alpha_l)$ and $\hat{\theta}^b(\alpha_u)$ of $\hat{\theta}^b$, where $\alpha_l = 0.025$ and $\alpha_u = 0.975$. In contrast the bias corrected confidence interval CI_{BC} uses the percentile $\hat{\theta}^b(\tilde{\alpha}_l)$ and $\hat{\theta}^b(\tilde{\alpha}_u)$ of $\hat{\theta}^b$ where

$$\tilde{\alpha}_l = \Phi(2\hat{z}_0 + z^{(\alpha)}) \text{ and } \tilde{\alpha}_u = \Phi\left(2\hat{z}_0 + z^{(1-\alpha)}\right) \qquad \text{Equation 3-7}$$

Φ is the standard normal cumulative distribution function, $z^{(\alpha)}$ is the α percentile of the standard normal distribution, and the value of the bias correction $\hat{z}_0$ is obtained directly from the proportion of bootstrap replications less than the original estimate $\hat{\theta}$,

$$\hat{z}_0 = \Phi^{-1}\left[\frac{Number\ of\ times\ that\ \left(\hat{\theta}^b < \hat{\theta}\right)}{B}\right] \quad \text{Equation 3-8}$$

The standard error of parameter $\hat{\theta}$ can be calculated as

$$\widehat{s.e.(\hat{\theta})} = \sqrt{\sum_{b=1}^{B} \frac{\left(\hat{\theta}^b - \overline{\hat{\theta}^b}\right)^2}{(B-1)}} \quad \text{Equation 3-9}$$

where

$$\overline{\hat{\theta}^b} = \sum_{b=1}^{B} \hat{\theta}^b / B \quad \text{Equation 3-10}$$

The bootstrap estimate of bias of the original estimate $\hat{\theta}$ is

$$\widehat{bias_{\hat{\theta}}} = \overline{\hat{\theta}^b} - \hat{\theta} \quad \text{Equation 3-11}$$

3.2.4.4 Bootstrap pivot approach for hypothesis testing

In terms of hypothesis testing, bootstrapping is often used as an alternative to inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors. When testing the null hypothesis that a scalar parameter is zero ($\theta = 0$), a significance test can be constructed from a confidence interval for said parameter: The null hypothesis is rejected at significance level 5% if the value is not included in the 95% bootstrap confidence interval (known as the *pivot approach* for constructing tests) (Davison and Hinkley, 1997). The traditional pivot approach can be used to calculate the p-value for testing the null hypothesis ($\theta = 0$) versus the alternative hypothesis ($\theta \neq 0$). The pivot for coefficient parameter θ is $\mathrm{T} = \Theta - \theta$, where Θ is an estimator of θ. Then the two sided p-value attached to the observed test statistic $t = \left(\hat{\theta} - 0\right) = \hat{\theta}$ is

$$p = \Pr(T^{*2} \geq t^2 | \hat{F}) \quad \text{Equation 3-12}$$

with $T^* = \hat{\theta}^b - \hat{\theta}$ and $\hat{F}$ is the bootstrap sampling distribution of $\hat{\theta}$. The pivot p-value is equivalent to the percentile confidence interval if the sampling distribution is symmetric.

3.3 A new MI-BT combined approach for mediation analysis in the presence of missing values

The purpose of this section is to establish a new approach that combines mediation analysis, Multiple Imputation and the bootstrap method. This new approach is named the MI-BT approach, where 'MI' emphasises the use of Multiple Imputation and 'BT' that of the bootstrap. The MI-BT method enables mediation analysis allowing observed confounders, in the presence of missing data and the hierarchical structures implied by the parenting programme trials. This section consists of four parts. The first part shows how to construct a point estimate for a causal mediation effect, the second part describes how to generate measures of precision of the estimator, the third part describes the implementation of the new MI-BT procedure and the last part evaluates the statistical properties of the approach.

3.3.1 Constructing the estimators of the causal mediation parameters

3.3.1.1 Mediation model and associated maximum likelihood type estimator

Under the MI-BT approach, the analysis models adjust the effects of measured confounders of the mediator – outcome relationship by conditioning on the confounders. The analysis models also account for the hierarchical data structure implied by the trial design by including random effects. Taking the SPOKES trial as an example, and assuming a list of known measured confounders, the mediation analysis models I have set up include three random effects to account for the three-level structure. Suppose we have n level-1 units (individual parents) nested within l level-2 units (therapy groups in the IY arm), which are nested within k level-3 units (school-year strata). In matrix notation,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\delta} + \mathbf{R}\gamma + \mathbf{M}\beta + \boldsymbol{\Psi}^{(3)}u^{(3)} + \boldsymbol{\Psi}^{(2)}u^{(2)} + \varepsilon_Y \qquad \text{Equation 3-13}$$

$$\mathbf{M} = \mathbf{X}\boldsymbol{\theta} + \mathbf{R}\alpha + \boldsymbol{\Psi}^{(3)}w^{(3)} + \boldsymbol{\Psi}^{(2)}w^{(2)} + \varepsilon_M \qquad \text{Equation 3-14}$$

where $\mathbf{Y}$ is $n \times 1$ vector of child outcome, $\mathbf{X}$ is an $n \times p$ matrix of known confounders, $\mathbf{R}$ is $n \times 1$ vector of randomised intervention indicator, $\mathbf{M}$ is $n \times 1$ vector of putative mediator. The vectors $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ present a set of coefficients indicating the strength of the relationship between confounders and child outcome, and confounders and mediator respectively. $u^{(3)}$ and $w^{(3)}$ are the level-3 school-year strata random effects for child outcome and mediator respectively, where $u^{(3)} = \left(u_1^{(3)}, u_2^{(3)}, \dots, u_k^{(3)}\right)^T$, $w^{(3)} = \left(w_1^{(3)}, w_2^{(3)}, \dots, w_k^{(3)}\right)^T$. $u^{(2)}$ and

$w^{(2)}$ are the level-2 therapy groups in the IY arm random effects for child outcome and mediator respectively, where $u^{(2)} = \left(u_1^{(2)}, u_2^{(2)}, \ldots, u_l^{(2)}\right)^T$, $w^{(2)} = \left(w_1^{(2)}, w_2^{(2)}, \ldots, w_l^{(2)}\right)^T$. Since the therapy groups are only assigned to the intervention arm, $u^{(2)}$ and $w^{(2)}$ are not available in the control arm. I fit a three-level mixed model with random intercept for both level-3 school-year strata and level-2 therapy groups in the treated arm. $\mathbf{\Psi}^{(3)}$ is a $n \times k$ design matrix for level-3 random effects $u^{(3)}$ and $w^{(3)}$, and is displayed below. $\mathbf{\Psi}^{(2)}$ is a $n \times l$ design matrix for level-2 random effects $u^{(2)}$ and $w^{(2)}$, and is displayed below. It is assumed that $u^{(3)}$, $u^{(2)}$ and ε_Y, are independent and follow an unspecified distribution with expectation equal to zero.

$$\mathbf{\Psi}^{(3)} = \begin{matrix} u_1^{(3)} & u_2^{(3)} & \ldots & u_k^{(3)} \end{matrix}$$
$$\begin{bmatrix} 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 1 & & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & & \ddots 0 \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \\ 0 & 0 & & 1 \end{bmatrix}$$
$$\begin{matrix} w_1^{(3)} & w_2^{(3)} & \ldots & w_k^{(3)} \end{matrix}$$

$$\mathbf{\Psi}^{(2)} = \begin{matrix} u_1^{(2)} & u_2^{(2)} & u_3^{(2)} & \ldots & u_l^{(2)} \end{matrix}$$
$$\begin{bmatrix} 1 & 0 & 0 & & 0 \\ 1 & 0 & 0 & & 0 \\ 1 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ 0 & 0 & 1 & & 0 \\ 0 & 0 & 1 & & 0 \\ 0 & 0 & 1 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & & \ddots 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & & 1 \\ 0 & 0 & 0 & & 1 \end{bmatrix}$$
$$\begin{matrix} w_1^{(2)} & w_2^{(2)} & w_3^{(2)} & \ldots & w_l^{(2)} \end{matrix}$$

Based on linear mixed models appropriate for the trial design, such as those described in Equation 3-13 and Equation 3-14 for SPOKES, conditional causal mediation effects (conditional on measured confounders $\mathbf{X}$) can then be estimated by constructing ML estimates of the fixed effects. ML estimation requires a fully parametric model and here the

formula for ML estimation under normality of the random effects and the residual terms is used. The causal effect of the treatment on the mediator (ETM) is estimated by $\hat{\alpha}_{ML}$; the effect of the mediator on the outcome (EMO) is estimated by $\hat{\beta}_{ML}$; the direct effect of treatment on outcome (DE) is estimated by $\hat{\gamma}_{ML}$; the indirect effect (IE) by $\hat{\alpha}_{ML}\hat{\beta}_{ML}$; and the total effect (TE) by $\hat{\gamma}_{ML} + \hat{\alpha}_{ML}\hat{\beta}_{ML}$. It is also worth pointing out that I use $\hat{\gamma}_{ML} + \hat{\alpha}_{ML}\hat{\beta}_{ML}$ to estimate the total effect of randomised treatment R on outcome Y in order to avoid obtaining an independent estimate of the total effect $\hat{\gamma}_{\text{total}}$ from a separate linear mixed model (regress Y on R). Although the ordinary least square estimators $\hat{\gamma} + \hat{\alpha}\hat{\beta}$ and $\hat{\gamma}_{total}$ are equal, the maximum likelihood estimators ($\hat{\gamma}_{ML} + \hat{\alpha}_{ML}\hat{\beta}_{ML}$) and $\hat{\gamma}_{total,ML}$ are only approximately equal. I would like to emphasise that the ML estimation is only used to construct point estimates of causal effects of interest. I am not exploiting normality-related properties of ML estimators such as standard errors of estimators provided by ML theory. However, as long as the random effect and residuals follow a symmetric distribution, one would expect the *ML-type* estimates to be approximately unbiased for large samples. Although no rigorous theoretical proof is provided in this project, I will later evaluate the bias of my proposed estimator empirically via bootstrapping (see Figure 3-5).

As outlined earlier, parenting trials are subject to missing values. I wish to use maximum likelihood estimation in this context due to its favourable properties under missingness (valid procedure under MAR). For the reasons summarized in Section 3.2.2, I will employ Multiple Imputation (MI) to construct an approximate ML-type estimator in the presence of missing values. Details of an MI procedure that can be employed to construct ML-type estimators of causal mediation effects in parenting trials will be given in the next section.

3.3.1.2 Multiple Imputation and the final MI-ML estimator

In this project, I use the flexible MI procedure *Multiple Imputation by Chained Equation* (MICE), as introduced in Section 3.2.2.2, to handle missing data. The imputation model is specified for each variable included in the imputation procedure. In a set of sequential regression multiple imputation models, linear regression is used for imputing normally distributed continuous variables, logistic regression is used for imputing binary variables, multinomial logistic regression is used for imputing nominal categorical variables and PMM is used for imputing the skewed continuous, count and discrete variables whose normality assumption is untenable.

A common question to MI is what should be included in the imputation model. In other words, how should the imputation model be specified? To avoid bias in the analysis model, the imputation model must include all variables that are in the analysis model (Robins and Wang, 2000). The variables included in the analysis model of parenting programme mediation analysis are child outcome, parenting mediators, randomised intervention group indicator and confounders of mediator-outcome relationship. Although the parenting mediators will be analysed one by one using a single mediator model, all putative mediators are included in the imputation model for the purpose of building up core imputation data that are valid for a set of mediation analyses proposed in this project.

You may recall that a specific feature of the IY parenting intervention trials is the hierarchical data structure (see Section 3.1.3). Taking the SPOKES trial as an example, it holds a three-level hierarchical data structure, i.e. individual participants (level 1) nested within therapy groups, which are nested within school-year strata (top level). Since the therapy groups are only available in the treated arm, the hierarchical structure is unbalanced. The therapy groups and school-year strata are modelled as random effects in the analysis step; however, they are modelled as fixed effects in the imputation step because as yet no tool is available to impute data under a three-level unbalanced random effect model. The most recent *REALCOM-IMPUTE* (Carpenter et al., 2011) software was developed for multilevel Multiple Imputation with mixed response types. However, it only handles two-level data and not three-level data. Comparing a fixed effects imputation model with a random effects model, the former makes fewer assumptions and uses more parameters, so that it is considered as a more general model. As discussed in Section 3.2.2.2, the literature (Fay, 1992, Schafer, 1999) supports the view that when the imputation model is more general than the analysis model, then MI leads to valid inferences. Therefore, modelling the trial design related variables as fixed effects in the imputation step is reasonable.

Further to the variables mentioned in the two paragraphs above, a list of auxiliary variables are also included in the imputation model for improving efficiency (using extra information) and reducing bias (predicted the missingness). In the SPOKES trial, the variables measured at baseline other than the selected mediator-outcome confounders, and the mediators and outcomes measured using different instruments other than the one included in the analysis model, are considered as auxiliary variables for the missing data imputation.

After setting up the imputation model, I will run the imputation multiple times (say H times) and generate H imputed data set. Then I will run the proposed analysis model (Equation 3-13 and Equation 3-14) using each set of imputed data and generate H set of ML-type estimators of the causal effects of interest. More specifically, the ML-type estimators are generated from each imputed sample, i.e. $\widehat{\alpha_\iota}$, $\widehat{\beta_\iota}$, $\widehat{\alpha_\iota}\widehat{\beta_\iota}$, $\widehat{\gamma_\iota}$ and $\widehat{\gamma_\iota} + \widehat{\alpha_\iota}\widehat{\beta_\iota}$ from imputation sample i. As reviewed in Section 3.2.2.2, the MI point estimator is simply the mean of the estimates calculated from H imputed dataset. The final approximate MI-ML estimators are simply the arithmetic average among the estimators of the total imputed ($H = h$) samples, i.e. $1/h\sum_1^h \widehat{\alpha_\iota}$, $1/h\sum_1^h \widehat{\beta_\iota}$, $1/h\sum_1^h \widehat{\alpha_\iota}\widehat{\beta_\iota}$, $1/h\sum_1^h \widehat{\gamma_\iota}$, and $1/h\sum_1^h(\widehat{\gamma_\iota} + \widehat{\alpha_\iota}\widehat{\beta_\iota})$ for the effect of treatment on mediator (ETM), effect of mediator on outcome (EMO), indirect effect (IE), direct effect (DE) and total effect (TE) respectively.

In summary, for these causal mediation effect estimators to be approximately unbiased, the following assumptions need to hold: 1) linearity of the relationships between the mediator and the clinical outcome and the covariates and the outcomes, 2) there is no unmeasured confounding of the M- Y relationship, 3) there is no $R \times M$ interaction effect on Y and 4) the missing data generating mechanism is MAR.

3.3.1.3 Nonparametric bootstrapping for generating statistical inference

The nonparametric bootstrapping method is utilised in the new MI-BT approach. My reasoning for using nonparametric bootstrapping to generate statistical inference has been provided in Section 3.2.4. For multi-level hierarchically structured data, the nonparametric cluster bootstrap method, sampling the entire case with replacement at the highest level from the observed data, will be used (see Section 3.2.4.2). In the context of trials, sampling at the highest level ensures that all design features that operate “below” this level are preserved in the bootstrap samples. For example, as described in Chapter 2, designs of parenting programme trials might utilise stratified randomisation or vary the randomisation ratio over time. Then, provided that the level of the variable that determines the randomisation is nested within (or is equal to) the level at which resampling occurs, such design features are maintained in the resample and thus the bootstrapping approach mimics the trial data generation mechanisms.

Taking the SPOKES trial as an example, the first step is to identify the trial design: As described in Chapter 2, the SPOKES trial is an RCT with two treatment arms. It employs randomisation within school-year strata and therapy groups are available only to those randomised to the active intervention group. The SPOKES trial design leads to a three-level hierarchical data structure: that is, individual participants are nested within the therapy groups, which are nested within school-years. The school-year strata (ten strata in total) are in fact the level at which the randomisation operates. To mimic this trial data generation procedure, I use a bootstrap sampling strategy: that is, resampling school-years (highest level) with replacement. In this bootstrap approach, the bootstrap resamples are drawn from the original observed data; the sampling unit is the whole school-year stratum; each bootstrap sample has the same number of school-years as the original observed data but potentially different numbers of parents. Besides, the original SPOKES trial sample contains missing values, and so do the bootstrap samples. In summary, this cluster bootstrapping method preserves the hierarchical data structure, missing data generating process and lower-level trial design features.

After replicating the bootstrapping B times, the procedure provides B bootstrap samples. Each has the same number of sampling units and might include missing values. Then MI-ML estimators are constructed for each bootstrapped sample by employing MI. This provides B final MI-ML point estimates for each causal mediation parameter of interest. Bias corrected BT confidence intervals and significance tests can then be calculated using the method described in Section 3.2.4.3 and the standard error of the estimate can be calculated based on Equation 3-9.

3.3.2 Implementation of MI-BT combined approach

3.3.2.1 Programming steps for implementing the MI-BT combined approach

Following the methodological description of the MI-BT approach in the previous section, I focus on the programming aspects of the combined approach in this section. A clear step-by-step guide is provided to construct the MI-BT point estimate and confidence interval of the causal parameters of interest.

Firstly, the point estimate of the parameters of interest is simply the MI-ML estimator calculated from the original data using MICE to handle the missing data and based on linear

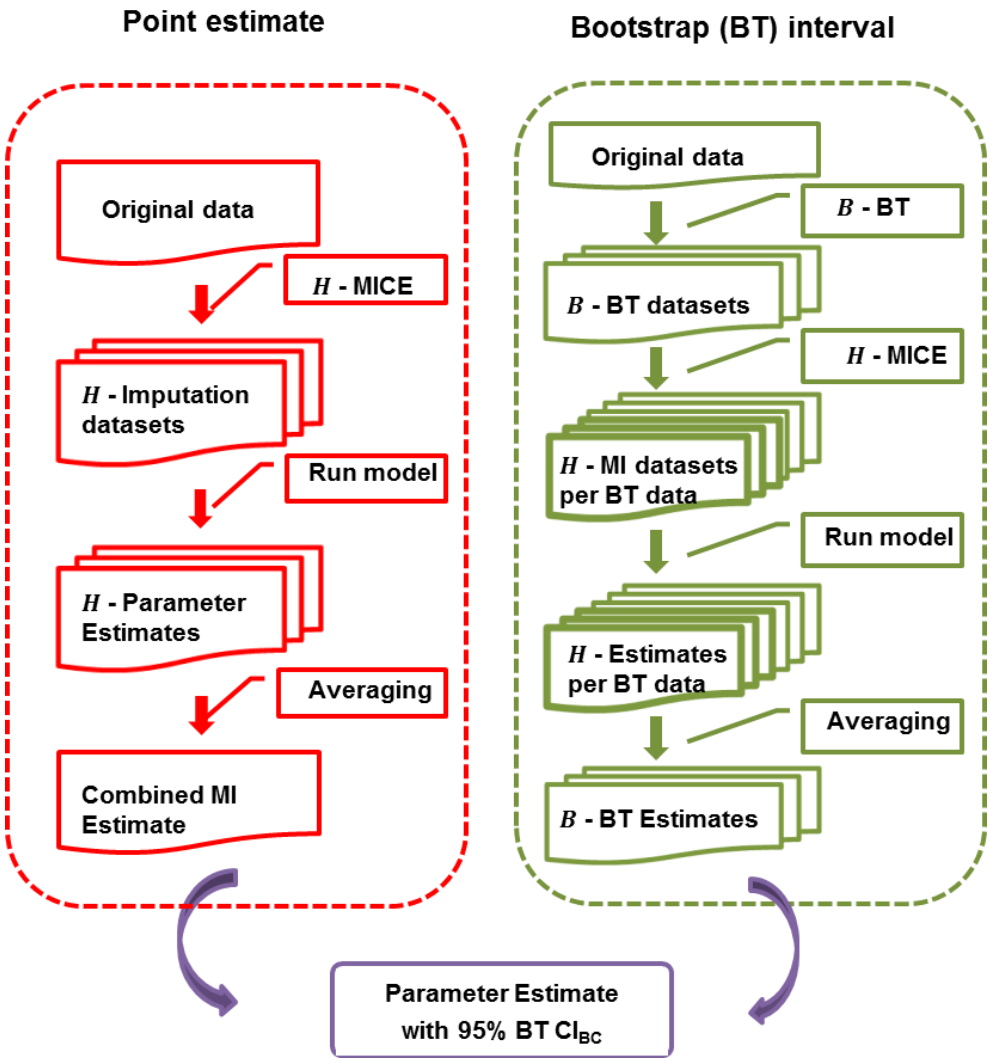
mixed effect mediation models to account for the hierarchical structure implied by the trial design. Details of how to construct the final MI-ML estimator have been provided in Section 3.3.1.2.

Secondly, generating the confidence interval of the causal parameters of interest using the MI-BT combined procedure involves four steps:

- **Step 1 – Resampling:** Consider the original dataset (with missing values) as a sample from a population that has a multi-level hierarchical structure with size of N for the highest level units (e.g. school-year strata in the SPOKES trial), and then draw B bootstrap samples of N units (school-year strata) at the highest level randomly with replacement from the original clustered dataset. B is the number of bootstrap samples. The bootstrap samples usually contain missing data as well.
- **Step 2 – Multiple Imputation:** For each bootstrap sample, run MICE and generate H imputed datasets, where H is the number of imputations. Hence, H imputed datasets are generated per bootstrap sample.
- **Step 3 – Construction of MI-ML estimators:** For each imputed dataset, fit the proposed linear mixed effect mediation models and estimate the parameters of interest. For each parameter of interest, this provides H estimates. The arithmetic average of the H estimates is the estimate of the parameter of interest derived from the particular bootstrap sample. B bootstrap samples lead to a set containing B estimates.
- **Step 4 – Bootstrap inferences**: The set of B bootstrap estimates forms the empirical distribution of the estimator of the parameter of interest. This empirical sampling distribution provides the basis for the construction of a confidence interval or significance test for the parameters of interest.

To provide an intuitive understanding of the procedure, Figure 3-2 illustrates how the procedure combines mediation, MICE and BT methods to construct the point estimate and associated inferences for the parameters of interest. The left block illustrates the construction of the point estimate from the original data sample and the right block illustrates the construction of an associated BT confidence interval. The number of imputations is H and the number of bootstrap resamples is B.

Figure 3-2 Mediation Analysis, MI-BT Combined Procedure



I have programmed this MI-BT combined procedure, as shown in Figure 3-2, in the general purpose statistical package Stata. An example annotated Stata code for illustrating the implementation of the MI-BT combined approach is listed in the Appendix II of this thesis. The Stata programme consists of two parts: part I generates the point estimate of the parameters of interest and part II constructs the associated confidence intervals. The confidence interval generating procedure is computationally expensive because the two iterative processes (BT and MI) require constructing MI for each BT sample and running a mediation model for each imputed data. Taking SPOKES as an example, constructing MI with $H = 20$ for $B = 1000$ bootstrap samples took 21 hours and 15 minutes, and running a single mediator mediation model using the $20{,}000$ $(B \times H)$ imputed data set took 6 hours. As mentioned in Section 3.3.1.2, the 2000 imputation data are valid for the sequence of mediation analyses proposed in this project, so that there is no need to repeat MI for individual single mediator mediation analysis. Therefore, I would say that 6 hours per single mediator mediation analysis is an acceptable program running duration.

3.3.2.2 Selecting measured confounders for inclusion in the analysis model

The MI-BT combined approach discussed above assumes that the investigator can specify the analysis model, i.e. that there is a list of known confounders of the mediator-outcome relationship. In practice, we may have a set of candidate confounders; it might not be feasible to include all of them in the model and there may be uncertainty as to which of them represent true confounders. Thus, in this case a confounder selection procedure is required before mediation modelling can commence. Again, I use the SPOKES trial as an example. This trial measured a rich set of potential confounders including child age, gender and ethnicity, mother's education and depression, family income, eligibility for free school meals, lone parent and family. Given the sample size ($N = 112$) of this trial, it is not practically feasible to condition on all these potential confounders in the mediation model. I propose two criteria for selecting a shortlist of the measured confounding variables ($\mathbf{X}$) to be included in the mediation analysis models:

1) Include the known confounding variables in the analysis models, such as the child outcome measured at baseline and the parenting practice mediator measured at baseline. More specifically, these mediator and outcome variables measured at baseline should predict their values measured at the end point respectively, and previous research with longitudinal data (MacKinnon, 2008) suggested that an earlier time mediator might be associated to a later time outcome and an earlier time outcome might be associated to a later time mediator. Therefore, the outcome and mediator measured at baseline are considered as "definite" confounders of the mediator-outcome relationship.
2) If one has a set of theoretically defined potential confounders which have been measured, then one can employ a procedure (as listed below) to empirically select a subset of variables for inclusion.
    - Assess the change of the effect of the mediator on the outcome ($\Delta\beta$) after adding the putative confounder to the outcome model (Equation 3-13) with variables randomisation (R), parenting mediator (M), child outcome measured at baseline, and parenting mediator measured at baseline already included in the model. $\Delta\beta = \beta_0 - \beta_c$, where β_0 and β_c are the effects of M and Y without and with test confounder respectively.

- Confidence limits for this change can be calculated using the new MI-BT combined procedure that was introduced in Section 3.3.1.3. The parameter to be tested is $\Delta\beta$.
- For single mediator models, the list of candidate confounders are tested one at a time for each putative mediator (the parenting practices listed in Chapter 2, Table 2-9). This means that each candidate confounder is tested for each putative mediator based on the change in the respective $\Delta\beta$ for that mediator.
- Potential confounding variables that significantly changed any effect of the putative mediator on the outcome at 70% level (i.e. the 70% confidence interval of $\Delta\beta$ does not include zero or p-value no more than 0.3) are included in all single mediator mediation analysis models.

Here, I used a liberal 70% confidence interval (30% type I error) instead of the conventional 95% confidence interval because basically this procedure is intended to generate a range for selecting a set of possible confounders rather than a formal hypothesis testing procedure of a non-zero parameter. A higher type I error rate is less restrictive and allows me to include more potential confounders into the mediation model. Although a selected confounder is not necessarily a true confounder, including it into the mediation model will not harm the analysis, given that the total number of predictors in the analysis model is within the limit specified in the next paragraph.

The purpose of the confounder selection step is to maintain the accuracy of mediation effect estimation. Here, I borrowed the rule of thumb in terms of the maximum number of predictors in multiple linear regression allowed by the sample size. Green (Green, 1991) suggested a $sample\ size\ requirement\ of\ N > 104 + number\ of\ predictors$ for testing individual predictors (assuming a medium-sized relationship). Therefore, in the mediation model, the total number of explanatory variables (including the selected confounders, the randomisation variable and the mediator) need to be within the limit.

3.3.2.3 Choice of the number of imputations

An important question when implementing MICE is how many imputations are needed to ensure the efficiency (i.e. true variance) of the estimates and the reproducibility (i.e. Monte Carlo error) of the statistics. In the case of using MI to handle missing data in regression

analysis, the rule of thumb is "the number of imputations should be similar to the percentage of cases that are incomplete" (Von Hippel, 2009, White et al., 2011). However, this criterion was based on the standard regression parameter estimate and it might not fit into the indirect effect estimate (product of two parameters) derived from mixed effect models. Besides, the MI-BT combined method is computationally expensive, especially when applied multiple times to analyse multiple single-mediator models, and I was hoping that a smaller number of imputations could still be sufficient.

I propose an experimental approach to determining a sufficient number (H) of imputations for the conventional number of bootstrap samples ($B = 1000$). The basic idea is to collect the analysis results (point estimate and standard error) of the MI-BT mediation analysis under different numbers of imputations ($H = 1\ to\ h_{max}$) and investigate the impact of the number of imputations on the point estimate and the standard error of an example mediation effect in a single-mediator model. Here, I focus on the mediation effect, as it is the crucial and unique component of mediation analysis. The number of imputations will be selected based on both graphical and numerical evidence, as explained in detail later. The selected number of Multiple Imputations (h) is expected to sufficiently reduce the Monte-Carlo error of the effect estimate and will be applied to all proposed single-mediator models using the MI-BT mediation analysis approach. To determine the number of MI in a graphical manner, I line plot the point estimates and the standard errors under different numbers of Multiple Imputations to show the empirical trend of the analysis results with different numbers of imputations (see Figure 3-4). When the lines are approximately flat, it indicates that increasing the number of imputations will not benefit the analysis results. Then the point where the line became flat determines the number $(H = h)$ of Multiple Imputations. Previous imputation experiences suggested that 100 imputations are generally more than enough imputations, and this can also be verified by the plots. According to this, the analysis results of 100 imputations can be used as a reference to evaluate the analysis results of the selected number of imputations. Numerically, we can calculate the relative difference of the analysis results between using the selected number of MI $(H = h)$ and the reference number of MI $(h_{max} = 100)$ in order to evaluate the performance of the h imputations. I consider less than 5% relative difference of the analysis results (point estimate and standard error) as the criterion for accepting the selected $(H = h)$ number of imputations.

The disadvantage of the experimental approach to finding sufficient number of Multiple Imputations is intensive computing. To save computational power consumption, the simulation extrapolation (SIMEX) method (Carroll et al., 2012) might be considered for future applications on finding sufficient number of MI. SIMEX is a general methodology originally developed for estimating and correcting bias due to measurement error via simulation. With additive measurement error, the measurement $W_i = X_i + U_i$, where X_i is a true measure of individual i and U_i is a normal random variable with variance σ_u^2. The SIMEX approach simulates M data sets with different levels of the measurement error added, each with successively larger measurement error variance, say $(1 + \zeta_m)\sigma_u^2$, where $0 = \zeta_1 < \zeta_2 < \cdots < \zeta_M$ are known (simulation step). Then estimates are obtained from each of the generated contaminated data sets (estimation step). Next, the simulation and estimation steps are repeated a large number of times, and the average value of the estimate for each level of contamination is calculated. These averages are plotted against the ζvalues and a regression technique (i.e. nonlinear least squares) is used to fit an extrapolate function to the averaged, error-contaminated estimates. The extrapolation to the ideal case of no measurement error ($\zeta = -1$) yields the SIMEX estimate. Similarly, in our case, the Monte-Carlo error of the estimate and the measurement error of the estimate are alike, and MI estimates from different numbers (h) of the MI are analogues of the estimates obtained from different levels (ζ) of the measurement error. Thus, we can use the regression technique to fit an extrapolation function to the MI estimates against h values. By doing this, not only can the sufficient number of MI be decided, but also the SIMEX estimation of variance can be yielded via extrapolating to the ideal number of MI without actually performing so many imputations. The proposed SIMEX application in reducing Monte-Carlo error of MI is a heuristic logical deduction and is still very much in its infancy. The algorithm and the detailed methodological development are still required. Most importantly, the methods described above for correcting the additive error of the estimate shares the simplicity and generality of regression calibration theoretically. The generality of the methods indicates that it is reasonable to consider that the simulated estimates against different numbers of MI follows the same extrapolation function when applying it to mediation analyses with different mediators.

3.3.2.4 Standardised effect measures

In the single mediator model for treatment effects in RCT, the meaning of the indirect effect is the effect of the randomised treatment on the outcome that is indirect, i.e. operates through the mediator (intermediate variable). The value of the indirect effect is more interpretable if the unit of measurement of the outcome involved is clear. This is not always the case for child outcome measures in parenting trials. For example, the Parental Account of Children's Symptoms (PACS) is actually the average of the frequency score and the severity score of children's behaviour with a range from 0 to 3 (See Chapter 1, Section 1.2.2), so that it is difficult to interpret a unit difference for this score. The standardised effect is defined as one that is not wedded to a particular measurement scale (Preacher and Kelley, 2011), so that the meaning of the standardised effect is consistent under different applications and benchmarks derived from the standardised effect measures can be applied to new set of scales. The standardisation approach applied in this project is the ratio of the indirect effect to the standard deviation of the outcome measured at baseline (MacKinnon, 2008). I prefer to use the standard deviation of the outcome measured at baseline to the standard deviation of the outcome measured at the endpoint because the pre-randomisation standard deviation is a better estimate of the population variability. The same logic is applied to standardise other estimates of causal effects of interest. The formulae for calculating the standardised effects are listed below:

$$\hat{\alpha}_{Standardised} = \hat{\alpha}/S_M \quad \text{Equation 3-15}$$

$$\hat{\beta}_{Standardised} = \hat{\beta} * S_M/S_Y \quad \text{Equation 3-16}$$

$$\hat{\gamma}_{Standardised} = \hat{\gamma}/S_Y \quad \text{Equation 3-17}$$

$$\hat{\alpha}\hat{\beta}_{Standardised} = \hat{\alpha}\hat{\beta}/S_Y \quad \text{Equation 3-18}$$

$$(\hat{\alpha}\hat{\beta} + \hat{\gamma})_{Standardised} = (\hat{\alpha}\hat{\beta} + \hat{\gamma})/S_Y \quad \text{Equation 3-19}$$

S_Y is the standard deviation calculated from the outcome measured at baseline and S_M is the standard deviation calculated from the mediator measured at baseline. In addition, the proportion of the total effect that is mediated, $\hat{\alpha}\hat{\beta}/(\hat{\gamma} + \hat{\alpha}\hat{\beta})$, is also a commonly used way to gauge the size of the mediation effect. Both the standardisation method and the proportion method are applied in this project.

3.3.3 The statistical properties of the MI-BT combined approach

Supposing that there are no missing values and the sample does not have a hierarchical structure, it has been shown that the biased-corrected confidence interval approach provides a more powerful 95% confidence interval for $\hat{\alpha}\hat{\beta}$ (MacKinnon et al., 2004). For error terms/latent variables with symmetric distributions, I would also expect the MI-ML estimate of each bootstrap sample (with missing values) to be asymptotically unbiased. More generally, the bias-corrected confidence intervals and associated significance tests should hold their confidence/significance levels according to standard bootstrap theory (Efron and Tibshirani, 1993). Thus I suggest accompanying confidence intervals and tests with an assessment of bias. The latter can be provided by comparing the mean of the empirical bootstrap distribution of an estimator with its sample value. The difference between these two values is an estimate of the direction and size of any bias.

The bootstrap estimate of bias is a method for measuring the accuracy of the estimator (see Equation 3-11). A small value of the ratio of estimated bias to SE ($\widehat{bias_{\hat{\theta}}}/\widehat{s.e.(\hat{\theta})}$) indicates that the estimator's performance (e.g. the estimate and standard error) is desirable. We are resigned to the fact that $\hat{\theta}$ is a variable estimator of θ, but usually we do not want the variability to be overwhelmingly on the low side or on the high side. As a rule of thumb, a bias of less than 0.25 standard errors can be ignored and this means that no correction is needed to $\hat{\theta}$, see page 128 of Efron and Tibshirani's book (Efron and Tibshirani, 1993).

Another interesting point of this MI-BT combined method is that the distribution of the indirect effect estimator constructed from the bootstrap samples is actually an empirical sampling distribution of the estimator. It can be used as an experimental device to check the normality assumption made by the traditional mediation approach when generating inferences of the indirect effect and by the Multiple Imputation when combining results for imputed samples. Skewedness or kurtosis of the bootstrap sampling distribution of the estimator supports the application of nonparametric inference.

3.4 Application of the MI-BT combined approach to SPOKES

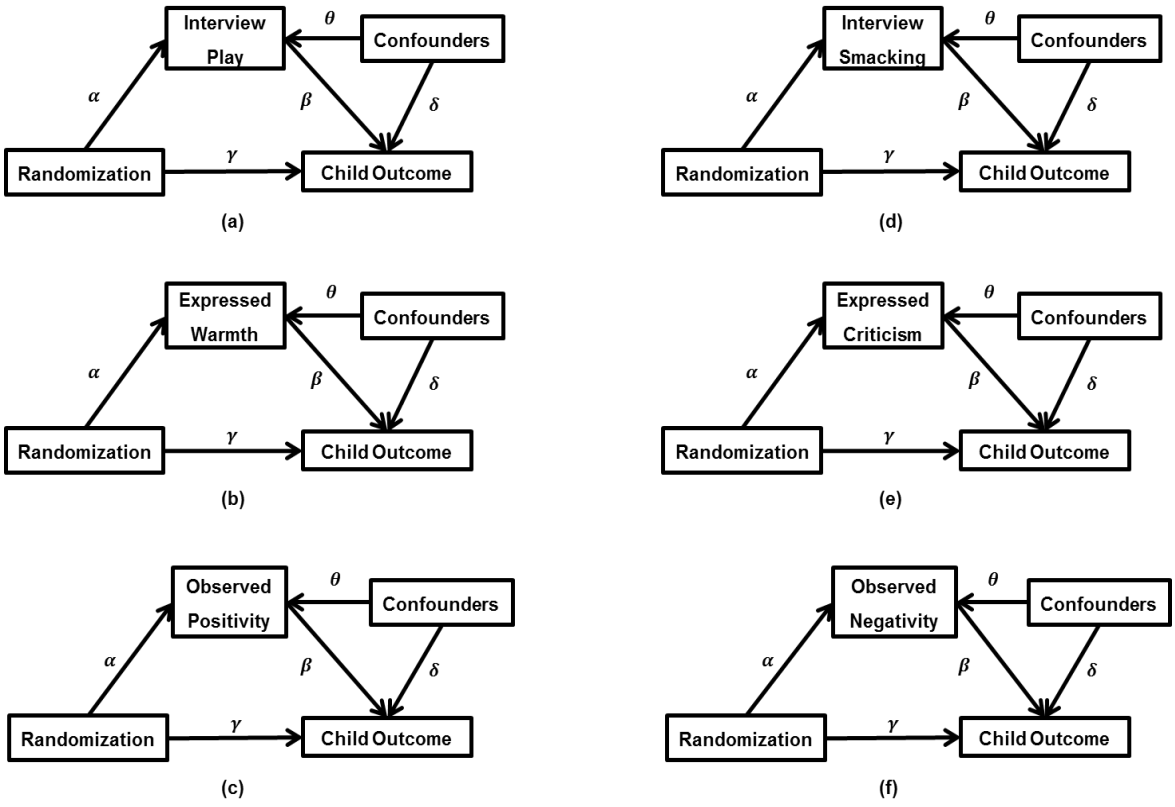
To illustrate the utility of the MI-BT combined mediation analysis approach with an actual psychology intervention trial, this method is applied to analyse data from the SPOKES "Incredible Years" (IY) Parent Training Intervention trial (Scott et al., 2010b). The SPOKES

trial is a randomised controlled trial of the IY parents' group training aimed at reducing children's antisocial behaviour. The total number of families participating is $n = 112$ with missing values appearing in potential baseline confounding variables, putative mediators and outcome variables. A detailed description of the SPOKES trial has been provided in Chapter 2 of this thesis.

3.4.1 The hypothesised single-mediator models of SPOKES

The putative mediators to be tested separately are six parenting practices that have been targeted by the intervention: *interview play, expressed warmth, observed positivity, interview smacking, expressed criticism,* and *observed negativity* (see Chapter 2 Table 2-9). Figure 3-3 shows the six hypothesised single-mediator models that need to be assessed using the proposed new mediation analysis method.

Figure 3-3 Single-Mediator Mediation Models for SPOKES trial

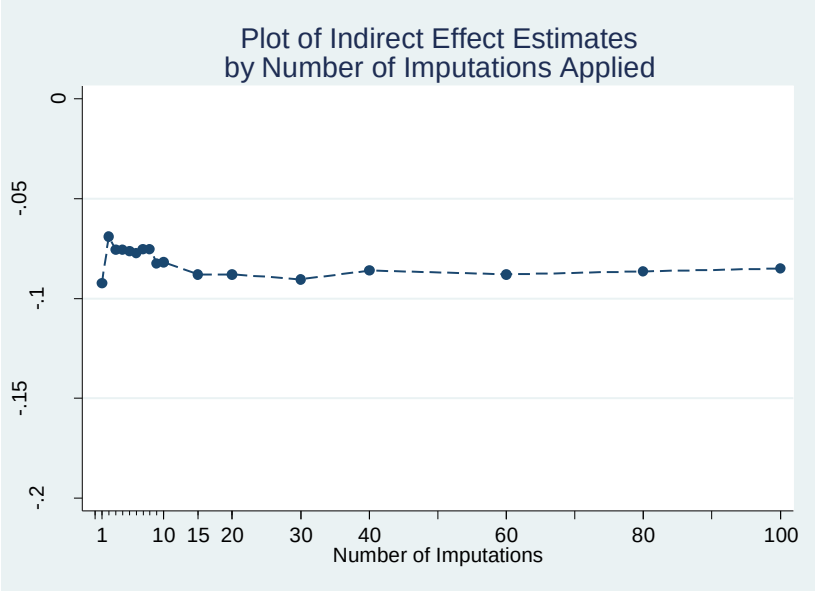


3.4.2 Deciding sufficient number of imputations

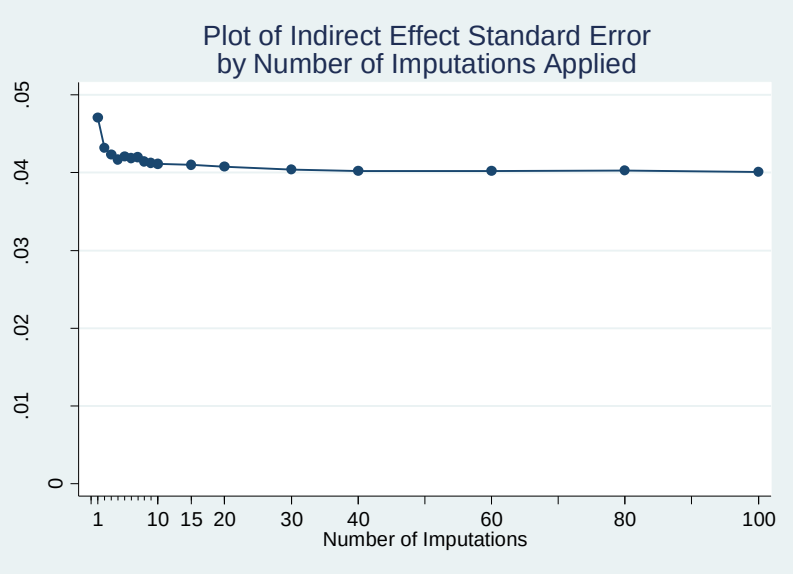
In the SPOKES trial, which has 50% missing data (50% incomplete cases), I run the mediation analysis for the putative mediator, negative expressed, using the proposed MI-BT procedure under different numbers of imputations including h = 1 to 10, 15, 20, 30, 40, 60, 80 and 100

with 1000 bootstrap simulations. The point estimates and standard errors of the indirect effect are plotted in Figure 3-4.

Figure 3-4 Indirect effect point estimates and standard errors under different number of imputations using SPOKES trial



(a)



(b)

The empirical results suggest that 20 imputations are sufficient because the relative difference of the point estimates between 20 imputations and 100 imputations is less than 5%, taking 100 imputations as reference, and the standard error inflation of 20 imputations

is less than 2% of 100 imputations. A detailed description of the procedure for determining the number of imputations has been provided in Section 3.3.2.3.

3.4.3 Selecting the confounders to be included in the mediation model

First of all, a set of baseline confounding variables of the mediator-outcome relationship needed to be selected from a pool of potential confounders using the selection approach described in Section 3.3.2.2 with 20 imputations and 1000 bootstrap. The key feature of this approach is that it accounts for both missing values and the hierarchical structure of the data during the confounder selection process. I proceeded as follows: For each single-mediator mediation model, child antisocial behaviour measured at baseline and parent mediator measured at baseline are included in the model as known confounders. For each single-mediator model, a list of candidate confounders, such as child's age, gender and reading ability, parent's education, ethnicity and depression, family size, lone parent and eligibility for free school meals, were assessed one by one. For each mediator and confounder, a "confounder test" was created by constructing a p-value for the hypothesis that there was zero change in the mediator effect ($\Delta\beta = 0$) after including the putative confounding variable in the analysis model. Since I would like to be liberal in picking up confounders, I made it easy for a test to detect a confounder by setting the type-1 error rate to 30% and did not adjust for multiplicity. The results of these tests are listed in Table 3-2. Five variables were found to act as confounders for at least one mediator-outcome relationship (shaded areas in Table 3-2). These variables are (1) child's gender, (2) child's reading ability, (3) parent's education, (4) parent's depression and (5) lone parent.

As a result, the matrix $\mathbf{X}$ in the models described in Equation 3-13 and Equation 3-14 contains seven baseline confounding variables: the five selected confounding variables plus the child outcome measured at baseline and the parenting practice mediator measured at baseline. Regression coefficients of these variables are included in respective models as fixed effects. In contrast, therapy groups in the treated arm $\Psi^{(2)}$, and the higher level school-year strata $\Psi^{(3)}$ are included in the models as random effects to account for the hierarchical structure of SPOKES data. In respective models, R denotes the randomly assigned IY parenting programme. Y is the child antisocial behaviour measured at one year after randomisation. M is one of the six putative mediators to be tested. Finally, the fixed effects of R and M in respective models represent our causal mediation parameters of interest.

Table 3-2 Results of testing confounders of mediator-outcome relationship

Confounders	Parameters	Δβ of Putative Mediators					
		Interview play	Expressed warmth	Observed positivity	Interview smacking	Expressed criticism	Observed negativity
Lone parent	Estimate	-0.003	0.004	-0.005	-0.001	0.044	0.033
	70% LowCI	-0.013	-0.007	-0.016	-0.013	0.021	0.028
	70% UpCI	0.003	0.013	0.005	0.013	0.071	0.051
	p-value	0.655	0.677	0.605	0.929	0.064	0.008
Eligibility for free school meals	Estimate	0.001	-0.006	0.002	0.001	0.007	-0.004
	70% LowCI	-0.002	-0.024	-0.004	-0.003	-0.006	-0.013
	70% UpCI	0.006	0.003	0.008	0.007	0.02	0.007
	p-value	0.813	0.594	0.663	0.818	0.552	0.665
Child reading ability	Estimate	0.007	0.007	0.004	-0.006	-0.002	0.032
	70% LowCI	0.004	-0.005	0	-0.017	-0.013	0.015
	70% UpCI	0.015	0.023	0.011	0.004	0.006	0.074
	p-value	0.205	0.587	0.402	0.55	0.798	0.153
Parent depression	Estimate	-0.003	-0.002	-0.004	0.003	0.015	0.001
	70% LowCI	-0.009	-0.007	-0.015	0	0.01	-0.001
	70% UpCI	-0.001	0	0	0.024	0.034	0.012
	p-value	0.307	0.392	0.347	0.38	0.079	0.806
Child age	Estimate	0.005	0	0.004	-0.001	0	0.002
	70% LowCI	-0.002	-0.009	-0.001	-0.013	-0.009	-0.002
	70% UpCI	0.018	0.007	0.017	0.002	0.014	0.027
	p-value	0.634	0.909	0.564	0.798	0.976	0.854
Child gender	Estimate	0.002	0.002	-0.005	0.002	0.014	0.005
	70% LowCI	-0.003	-0.001	-0.012	-0.003	0.002	-0.001
	70% UpCI	0.007	0.007	-0.001	0.007	0.032	0.015
	p-value	0.733	0.584	0.281	0.733	0.293	0.484
Parent ethnicity	Estimate	-0.002	-0.002	-0.001	0.003	0.001	-0.006
	70% LowCI	-0.015	-0.013	-0.006	0	-0.004	-0.026
	70% UpCI	0.001	0.001	0	0.018	0.008	-0.003
	p-value	0.666	0.716	0.746	0.554	0.844	0.434
Parent education	Estimate	0	-0.001	0.001	0	0.022	0
	70% LowCI	-0.003	-0.012	-0.003	-0.004	0.005	-0.004
	70% UpCI	0.003	0.001	0.008	0.009	0.04	0.009
	p-value	0.995	0.814	0.89	0.944	0.223	0.944

Note: The shaded areas indicate significant change of mediator-outcome relationship (Δβ) at 70% significant level.

3.4.4 Mediation analysis using the MI-BT combined approach

After setting up the mediation analysis models for SPOKES, six putative mediators in the single-mediator models displayed in Figure 3-3 are tested separately following the MI-BT combined procedure described in Section 3.3. For SPOKES, the highest clustering level is school-year. Thus, the cluster bootstrap was applied at this level. Missing values were handled using MICE, within which the PMM technique was employed to ensure that imputed values were plausible for variables with discrete distributions such as "expressed warmth" or "expressed criticism".

For each single-mediator model, the effect of the IY parenting intervention on the mediator ETM (α), the effect of the mediator on the child outcome EMO (β), and the direct effect DE (γ), the indirect effect IE ($\alpha\beta$) and the total effect TE ($\alpha\beta + \gamma$) of the IY parenting intervention on child outcome conditioning on a set of measured baseline confounding variables (X) are estimated with an associated 95% bias corrected bootstrap confidence interval. The calculation of the estimates takes into account the missing values and the hierarchical data structures, and it also relaxes the normality assumption of the traditional Baron and Kenny type mediation analysis. The estimate of each mediation effect is standardised by the standard deviation of the corresponding measurement at baseline. The pivot method based two sided p-values are calculated, assuming that the sampling distribution of the parameter estimate is symmetric. However, the approximate bootstrap sampling distributions may be skewed or have a high kurtosis value, and therefore the p-value, standard error and the bias corrected confidence interval might not entirely agree with each other. The results of a mediation analysis for each putative mediator using the new MI-BT method are shown in Table 3-3.

Table 3-3 Results of mediation analysis for each putative mediator using the MI-BT method

Putative Mediator	Causal mediation parameter	Estimate	SE	P-value	Bias Corrected 95% BT CI
Interview Play	α	0.40	0.18	0.03	(0.03, 0.73)
	β	0.11	0.11	0.26	(-0.05, 0.39)
	γ	-0.54	0.12	<0.01	(-0.72, -0.21)
	$\alpha\beta$	0.04	0.06	0.34	(-0.01, 0.26)
	$\gamma + \alpha\beta$	-0.50	0.13	<0.01	(-0.68, -0.05)
Expressed Warmth	α	0.46	0.31	0.14	(-0.05, 1.21)
	β	-0.24	0.09	0.01	(-0.41, -0.07)
	γ	-0.39	0.15	0.01	(-0.61, 0.04)
	$\alpha\beta$	-0.11	0.06	0.08	(-0.25, -0.01)
	$\gamma + \alpha\beta$	-0.50	0.13	<0.01	(-0.7, -0.14)
Observed Positivity	α	0.62	0.47	0.19	(-0.16, 1.68)
	β	0.01	0.07	0.90	(-0.09, 0.2)
	γ	-0.52	0.13	<0.01	(-0.71, -0.18)
	$\alpha\beta$	0.01	0.05	0.82	(-0.05, 0.24)
	$\gamma + \alpha\beta$	-0.51	0.13	<0.01	(-0.69, -0.14)
Interview Smacking	α	-0.32	0.10	<0.01	(-0.53, -0.15)
	β	0.17	0.12	0.17	(-0.02, 0.45)
	γ	-0.46	0.15	0.01	(-0.69, -0.06)
	$\alpha\beta$	-0.05	0.04	0.19	(-0.17, 0)
	$\gamma + \alpha\beta$	-0.51	0.13	<0.01	(-0.7, -0.13)
Expressed Criticism	α	-0.44	0.21	0.04	(-0.86, -0.03)
	β	0.43	0.10	<0.01	(0.26, 0.68)
	γ	-0.31	0.16	0.04	(-0.54, 0.13)
	$\alpha\beta$	-0.19	0.09	0.05	(-0.42, -0.05)
	$\gamma + \alpha\beta$	-0.50	0.12	<0.01	(-0.68, -0.15)
Observed Negativity	α	-0.17	0.18	0.36	(-0.52, 0.18)
	β	0.02	0.11	0.81	(-0.12, 0.36)
	γ	-0.51	0.12	<0.01	(-0.68, -0.2)
	$\alpha\beta$	0	0.04	0.82	(-0.09, 0.05)
	$\gamma + \alpha\beta$	-0.51	0.11	<0.01	(-0.69, -0.2)

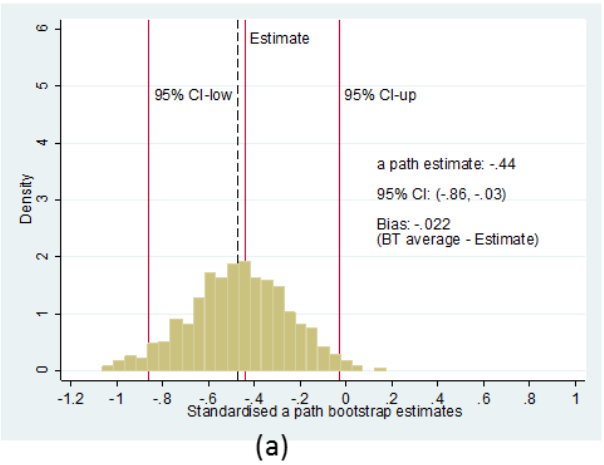
3.4.5 Interpretation of SPOKES mediation analysis results

Based on the 95% bias corrected bootstrap confidence interval of the IE ($\hat{\alpha}\hat{\beta}$), I found two mediators, parental expressed warmth and parental expressed criticism, of the TE ($\hat{\gamma} + \hat{\alpha}\hat{\beta}$) of the IY parenting intervention on reducing child antisocial behaviour conditioning on measured confounders. The results (see Table 3-3) show that 38% of the TE ($\hat{\gamma} + \hat{\alpha}\hat{\beta}$) is mediated via reducing parental criticism. The standardised indirect effect ($\hat{\alpha}\hat{\beta}$) of parental criticism is significant at 95% level with the estimates of -0.19 (-0.42, -0.05). The IY parenting intervention significantly reduced parental criticism ($\hat{\alpha}$) by -0.44 (-0.86, -0.03) and per unit

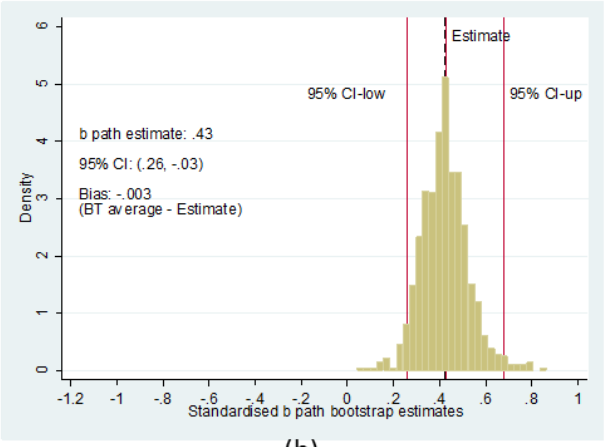
reduction of parental criticism leads to a significant reduction in child antisocial behaviour ($\hat{\beta}$) by 0.43 (0.26, 0.68). Independently, 22% of the TE ($\hat{\gamma} + \hat{\alpha}\hat{\beta}$) is mediated via increasing parental warmth. The standardised IE ($\hat{\alpha}\hat{\beta}$) of parental warmth is significant at 95% level with an estimate of -0.11 (-0.25, -0.01). There is a significant effect of the parental expressed warmth on reducing child antisocial behaviour with a standardised effect estimate ($\hat{\beta}$) of -0.24 (-0.41, -0.07). The effect estimate ($\hat{\alpha}$) of the IY parenting intervention on parental expressed warmth is 0.46 (-0.05, 1.21). Although it is not significant at 95% level, the size of the effect is relatively large, with a wide confidence interval. The above results provided empirical evidence for the mediation mechanism that the IY and literacy combined parenting intervention reduced child antisocial behaviour via improving parental expressed warmth and reducing parental expressed criticism, which adjusted the confounding effects of selected baseline factors (baseline child behaviour, baseline mediator, lone parent, child reading ability, child age, parental depression and parental education) and assumed that there is no unmeasured confounding between parenting mediator and child outcome.

To illustrate the results in an intuitive way, I draw a histogram (Figure 3-5) of the bootstrap estimates for each parameter, taking the mediator expressed criticism as an example. The effect estimate (red solid line), corresponding 95% bias corrected bootstrap confidence limits (red solid line), and the average value of the bootstrap estimates (black dashed line) are drawn as reference lines in the figures. The estimated biases of the estimators (see Equation 3-11) are represented by the differences between the point estimates calculated from the original data and the average of the bootstrap estimates. I find that bias values are less than 25% of the corresponding standard errors: the size of the estimated biases is therefore consistent with chance, and I therefore conclude that my parameter estimates are accurate (unbiased). In addition, the asymmetric bootstrap sampling distribution of the mediation effects estimates advocates the application of the nonparametric bootstrap approach to mediation analysis.

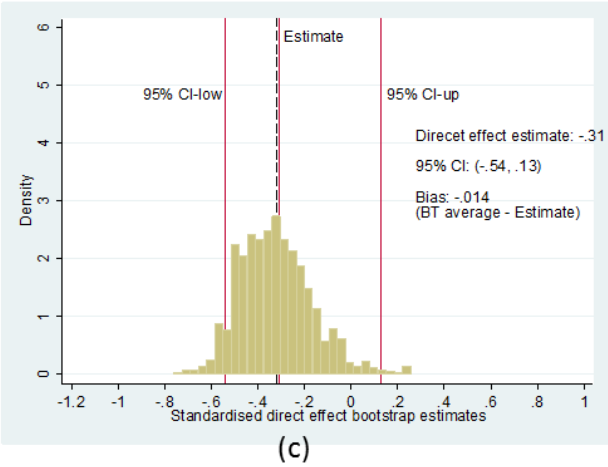
Figure 3-5 Histograms of BT estimates for causal parameter of interest when investigating mediation by expressed criticism



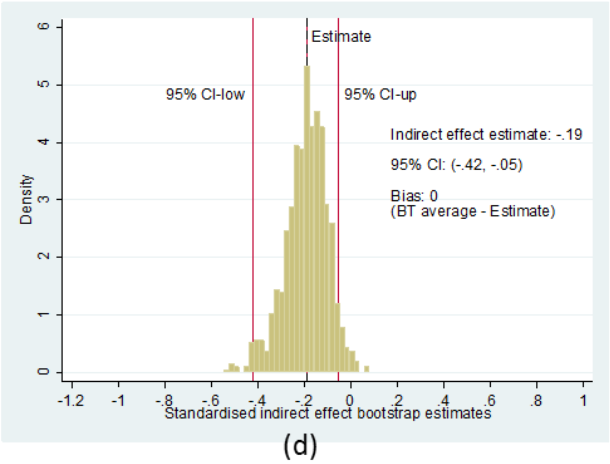
(a)



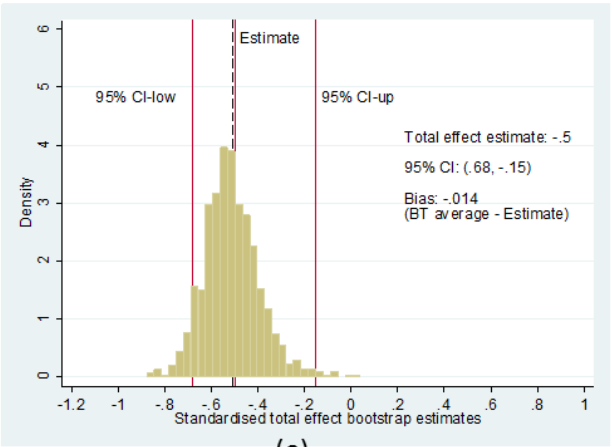
(b)



(c)



(d)



(e)

Note:
(a): Effect of treatment on expressed criticism
(b): Effect of expressed criticism on outcome
(c): Direct effect of treatment on outcome
(d): Indirect effect mediated by expressed criticism
(e): Total effect of treatment on outcome

3.5 Discussion

3.5.1 High standard of parenting intervention RCT

As introduced in Chapter 2, SPOKES is a rigorous randomised controlled trial with good quality measures using reliable measurement methods. The performance of randomisation leads to a causal interpretation of the total effect of intervention on the child outcome (TE)

and the effect of intervention on the parenting mediator (ETM). Benefiting from randomisation, the causal mediation analysis development can focus on resolving the confounding issue of the mediator-outcome relationship and does not need to consider the potential confounding bias in estimating the effects of intervention (TE and ETM). The trial measured multiple separate aspects of parenting behaviours, which allow us to investigate comprehensive multiple parenting mediators in one study. The applied measurement methods are reliable and standard, so that our assumption of accurate measurement is plausible. It is important to be clear that the total percentage of missing values is low. Although only 50% of the cases in the data are complete cases considering all mediators, outcomes and a list of baseline covariates, the percentage of missing values of individual variable is actually very low. This indicates that the data is missing sporadically among variables. The high percentage of non-complete cases for specific analyses is caused by including multiple variables with scattered missing values.

3.5.2 Why MI for each bootstrapped sample, instead of bootstrap for each imputed data?

This question was posed when I set up the procedure of combining MI and BT approaches for mediation analysis. It would be possible to first generate a set of imputed samples and then generate bootstrap inferences for each. This would yield a set of inferences: one for each imputed data set. These results would then have to be combined across imputations in some way to provide final results. However, there is no existing theory for combining multiple bootstrap variance estimates calculated from imputed data sets in order to construct the final variance estimate. Rubin's rules for combining multiple variances in MI were developed for maximum likelihood estimators. It is unclear how such an approach would perform when applied to bootstrap inferences. In contrast, the application of MI to each bootstrap sample to generate approximate ML estimators has a solid theoretical foundation. The bootstrapping ($H = 1000$) from the original data mimics the data generating procedure. This bootstrapping step is equivalent to running 1000 virtual trials. Then the model parameter is estimated from each virtual trial following the standard estimation procedure of Multiple Imputations and generates 1000 point estimates. The 1000 estimates are used to calculate the final estimate and associated confidence interval based on bootstrap theory. Therefore, the statistical inference based on the approach proposed in this project (MI for each bootstrap sample) is theoretical appealing and easy to understand.

3.5.3 Why use multiple imputations instead of deterministic or single imputation?

In the proposed MI-BT procedure, the multiple imputations are used purely to remove the variability around the predictions of the missing values by averaging; Rubin's rules are not employed to generate standard errors. Two questions are raised: firstly, since the bootstrapping has already taken into account the uncertainty of missing value predictions, why not use deterministic imputation to predict missing values, and secondly, why should we use multiple imputation instead of a single imputation?

Question 1: why not deterministic regression imputation? The answer is that the deterministic regression imputation may lead to biased estimate in mediation analysis. As we know, imputation needs to be conducted for both outcomes and covariates. If the source of added uncertainty (i.e. due to uncertainty coefficients in the imputation model and the random error in the regression model) is removed from the imputation, it would be biased for imputing covariates but unbiased for imputing outcomes. Deterministic imputation (there is no degree of randomness in the imputation process) of covariates must NOT use outcomes in the imputation model (Little, 1992), otherwise the associations between covariates and outcomes are exaggerated. The mediators took the roles of both outcomes and covariates at different times, so that deterministic imputation is biased in handling missing values for mediation analysis.

Question 2: why not single imputation?

In our case, the number of imputations (one or more) simply determines the size of the Monte-Carlo error in the variance of the estimator. Thus, the argument is simply about the precision of the estimator. For an infinite number of imputations, this Monte Carlo error is zero, while for $H = 1$ imputation, it is maximal. The empirical distribution of the standard error of an estimator following the increasing of the number of imputations has been shown in Figure 3-4 (b), which can be considered as empirical and intuitive support of the application of Multiple Imputations. In fact, this empirical approach found that the sufficient number of imputations is not too large with $h = 20$ imputations lead to reasonable precision. To put it briefly, imputing missing values multiple times depreciates the Monte-Carlo error in the variance of the estimator. Therefore, MI is applied for imputing missing values in mediation analysis, while accounting for the uncertainty of missing values in both covariates and outcomes and providing a precise estimator.

3.5.4 Comparison between the proposed procedure and the *bmen* procedure

The idea of combining bootstrap and MI in mediation analysis for dealing with missing data is not brand new. In a recently published paper (Zhang and Wang, 2013), a similar Multiple Imputation and bootstrap combined approach was proposed for mediation analysis with missing data. An R package named *bmen*[1] was developed for implementing this approach; I use bmen to refer to their approach. Basically, the bmen method is an SEM mediation analysis using an MI and nonparametric BT combined approach. The original data was resampled to generate bootstrap samples and SEM mediation analysis in combination with MI applied for each BT data. In this section, I will discuss the differences between my MI-BT procedure and the bmen procedure.

Firstly, the approaches differ in respect of the mediation analysis models employed. My MI-BT mediation analysis approach employs linear mixed models to account for hierarchical data structures such as those implied by trials design for testing parenting interventions. In contrast, the bmen approach assumes an SEM and the *bmen* package is built on the R package *sem* (Fox, 2006). The *sem* package contains functions for fitting general linear structural equation models (with observed and unobserved variables) using the RAM approach (McArdle and McDonald, 1984), and for fitting structural equations in observed-variable models by two-stage least squares (an instrumental variables approach, to be covered in the next Chapter). However, as far as I am aware, *sem* is not capable of handling a hierarchical data structure. Secondly, the methods differ in their approach to Multiple Imputation/normality assumptions. The MI approach applied in my MI-BT procedure is the relatively new MICE method, which can deal with different types of variables with missing values, whereas bmen's traditional MI approach is the traditional MI approach, which assumes that the variables in the imputation model arise from a multivariate normal distribution. In fact, the putative parenting mediators are discrete and not normally distributed in parenting programme trials. Therefore the multivariate normal distribution assumption made by SEM and traditional MI is unlikely to hold here. Finally, the approaches differ in their bootstrapping strategy. Although both approaches involve non-parametric bootstrapping, the bootstrapping strategies are not the same. My MI-BT approach employs cluster bootstrapping to mimic the data generation mechanism and reflect the clustered

[1] http://cran.r-project.org/web/packages/sem/sem.pdf

data structure. In contrast, the bmen bootstrapping assumes that all the cases are independent and no feature was added that might lead to potentially correlated samples. In summary, my MI-BT approach shows advantages in handling hierarchical data structures and accurately imputing missing values for non-normally distributed variables compared to the bmen approach.

3.5.5 The strengths of the proposed new MI-BT combined mediation analysis method

The MI-BT combined mediation analysis is a practical and flexible method for addressing measured confounding of mediation analysis in the presence of missing data. The MI-BT is considered as a practical approach for several reasons: 1) All measured confounding variables of the mediator-outcome relationship are included in the mediation model for calculating the conditional causal mediation effects of interest. This simple yet important step improved the traditional Baron and Kenny mediation analysis by making the no unmeasured confounding assumption more plausible and consequently reduced bias. 2) Missing data are commonly seen in mediation analysis, but the default method in most standard statistical software is complete case analysis. Including missing data imputation in mediation analysis can improve efficiency and reduce bias for mediation analysis. 3) MICE is a very well developed approach for imputing missing values. It allows the inclusion of auxiliary variables to increase the power of detecting mediation effects under MAR. Practically, well-performing Multiple Imputation statistical packages are available in standard software, such as *ice* in STATA and *mice* in R. 4) The nonparametric bootstrap approach relaxes the distributional assumptions for the variables included in the analysis model. In this way, the MI-BT approach allows different types (continuous, discrete or binary) of mediators and outcomes in the mediation model under the assumption of linearity. Meanwhile, bootstrapping is easy to understand and implement once the bootstrapping strategy is decided.

Moreover, the MI-BT is a flexible approach because of the following features: 1) The mediation models can reflect different types of trial design by modifying the structure of the random effect components in the mixed effect models. In the SPOKES trial, the random effects structure appropriately modelled the three-level hierarchical data structure. Following the rules of multilevel modelling, one can modify the structure based on the trial design. 2) The Multiple Imputation tool (MICE) used in this project is very flexible. It can deal

with non-normally distributed variables with missing values and most importantly it can impute variables that are functions of other variables, including interaction variable and other second-order variables. Potentially, it expands the range of types of variable that can be included in the mediation analysis model. 3) Although the mediation effects estimator was constructed by fitting two linear models in this project, the MI-BT approach could also be used in conjunction with causal effects estimators if they were constructed with known favourable properties. For example, the Stata packages *mediation* (Imai et al., 2010) and *paramed* (Emsley and Liu, 2013) allow the existence of interaction between the mediator and the randomisation and binary mediator. Then the MI-BT approach will be extended to deal with non-linear mediation analysis in the presence of missing data.

3.5.6 Limitations of the MI-BT method

The MI-BT method described in the current chapter provides unbiased causal mediation effect estimates for randomised controlled trials under the assumption of no unmeasured confounding of the mediator-outcome association after adjusting the measured confounders. However, some of the confounding variables might not be measured for practical reasons or simply due to lack of knowledge about the confounders in a new research area. Consequently, the estimate of the indirect effect of the randomised intervention on the outcome via the mediator might still be biased and the bias can be in two directions, meaning over- or underestimation of the effects of interest, or, say, getting artificial positive or negative mediation findings. This leads to the focus of the next chapter, which is developing the MI-BT method to relax the assumption of unmeasured confounding of mediator-outcome association and provide unbiased estimates of the cause mediation effects for RCTs in the presence of missing data.

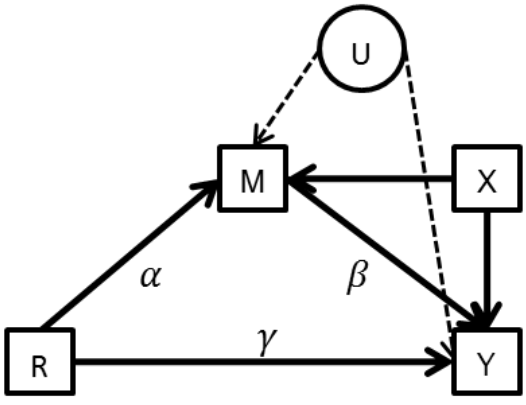
Chapter 4 Mediation Analysis Accounting for Unmeasured Confounding in the Presence of Missing Values

4.1 Introduction

4.1.1 Unobserved confounding in trials

In the previous chapter, the causal direct and indirect treatment effects (DE and IE) were estimated using the MI-BT approach under the assumption of no *unmeasured confounding* between the mediator and the outcome of the RCT. In other words, we assumed that all the confounders were measured and their impact was correctly modelled by our analysis models. However, unobserved confounding is a well-known threat to valid causal inference, and can seldom be ruled out with certainty if randomisation is not involved to investigate the causal effect of one variable on the other (Rubin, 1980). In the case of traditional mediation analysis of RCTs, the OLS estimate of the causal effect (β) of the post-randomisation mediator (M) on the outcome (Y) may still be subject to unobserved confounding bias upon conditioning on observed confounders (MacKinnon, 2008). The causal indirect effect ($\alpha\beta$) and the causal direct effect (γ) are not identified due to unmeasured confounding (U). Figure 4-1 illustrates a RCT single-mediator mediation model with both measured and unmeasured confounders. In this chapter, I go beyond the use of observed confounders to adjust confounding bias and propose to incorporate *instrumental variables* (Z) into the MI-BT approach to relax the assumption of no unmeasured confounding in mediation analysis.

Figure 4-1 RCT single-mediator model including unmeasured confounders

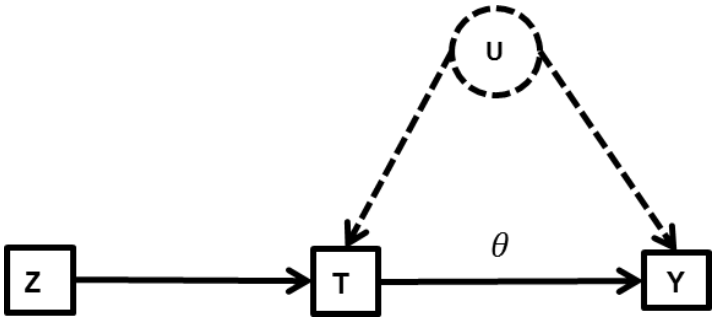


where R is the randomised treatment, Y is the outcome, M is the endogenous mediator, $\mathbf{X}$ is a set of exogenous measured mediator-outcome confounders and U is an unmeasured mediator-outcome confounder.

4.1.2 Instrumental Variables and their application to mediation analysis

The Instrumental Variable (IV) approach is widely used in the field of econometrics to overcome *endogenous explanatory variable* problems (Wooldridge, 2002). A variable is said to be endogenous when there is a correlation between the variable and the error term in a causal linear model. In contrast, an *exogenous variable* is independent of the error term. Standard least squares theory relies on exogeneity to provide unbiased estimates of causal effects and thus the OLS of a regression parameter is biased if the respective explanatory variable is endogenous. *Endogeneity* may arise when *unmeasured confounding* exists between the explanatory variable whose effect is of interest and the response variable. An IV is a variable that (i) explains part of the variability in the endogenous explanatory variable, and (ii) is uncorrelated with the error term of the linear model. As shown in Figure 4-2, T is the explanatory variable whose causal effect (θ) is of interest, Y is the outcome, U is an unobserved common cause of T and Y, and Z is an instrumental variable for T.

Figure 4-2 Instrumental variable

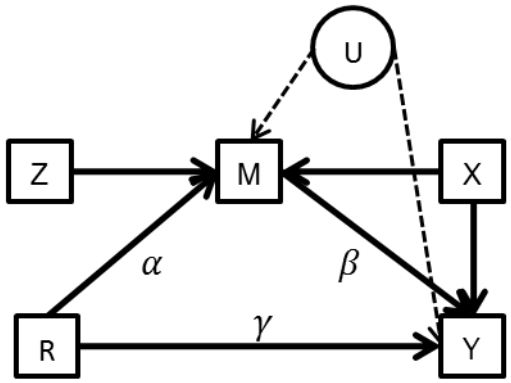


A general definition of an instrument Z is that it must be associated with the explanatory variable T; it must not share any common causes with the outcome; and it must not affect the outcome Y except through its effect on the explanatory variable T (so called *exclusion restriction*).

In the context of mediation investigations in clinical trials, even after including measured confounders in the analysis models for the mediator and outcome, the post-randomisation mediator might still be an endogenous explanatory variable in the outcome model due to remaining unmeasured confounding of the mediator-outcome relationship. The causal effect of the mediator on the outcome is not identified in the presence of unmeasured confounding. In order to correct the unmeasured confounding bias (endogeneity bias), I

propose to further develop the MI-BT mediation analysis approach using instrumental variables. Such instrumental variables (Z) must be associated with the mediator (M), but they are not associated with the error term in the regression model of the outcome (see Figure 4-3).

Figure 4-3 RCT single-mediator IV mediation model including unmeasured confounders



A crucial point of this approach is the existence of valid instruments. A number of suggestions have been made in the literature for *post hoc* IV selection: The most promising suggestion is the use of interactions between randomisation and baseline variables as instruments (Albert, 2008, Small, 2012, Emsley et al., 2010, Ten Have et al., 2007). The baseline variables influence the size of the effect of the treatment on the putative mediator (i.e. the moderators of the effect of treatment on the putative mediator), but they do not influence the size of the direct effect of the treatment on the outcome. It is important to recognise that the IVs assumptions are made for the interaction terms, but not for the main effect of the baseline variables on the mediator and the outcome. Gennetian et al. (Gennetian et al., 2005) have discussed the interaction between randomised intervention and baseline covariates, such as site in multisite randomized experiments and baseline characteristics such as age or gender, which might be valid instrumental variables. Emsley et al. (Emsley et al., 2010) have used two psychological treatment trials as examples to illustrate the utilisation of baseline covariates by randomisation interactions as IVs for estimating the direct and indirect effects. Small (Small, 2012) has discussed the assumptions when using baseline covariates that interact with random assignment as IVs in the setting that there is variation in effects across subjects and developed a method of sensitivity analysis for violations of the key assumption. Dunn et al. (Dunn et al., 2013) used the treatment by predictive biomarker (gene) interaction as an IV in the context of mediation analysis for stratified medicine.

Inspired by the applications of the IV approaches, this project aims to use randomisation by covariates interactions as IVs and incorporate the IV approach into the MI-BT mediation approach to generate the causal direct and indirect effects of the randomised parenting intervention (R) on the child behaviour outcome (Y) allowing for unobserved confounding (U) between the mediator (M) and the outcome (Y). The key ideas of using these interaction terms as IVs are: Firstly, randomisation ensures that there is no unmeasured confounding for the interaction instrument and the outcome, which offers a promising start as instruments. Secondly, it is assumed that the interaction effect operates solely by changing the mediator; hence the moderation is fully mediated. Thirdly, baseline values of the mediator or clinical outcome (severity) are good candidates. Aspects of treatment, such as parenting training therapy groups in this case, can also be used as IVs (Gennetian et al., 2002) because they are effectively the interactions between randomisation and process variables. However, as the effects of the treatment process variables in the control group are unobserved, using such variables as IVs requires a stronger assumption. Mendelian randomisation has been proposed as an IV approach to causal analysis (Smith and Ebrahim, 2003). As an analogy to randomization in a clinical trial, genetic markers are used as instrumental variables to estimate the causal relationship between a phenotype and an outcome variable. However, genetic information was not collected in the current parenting programme trials. Additionally, it is becoming clear that IV generation by design might be required (Dunn et al., 2013), though I am not pursuing this here, since my project is focussed on secondary analyses of existing parenting trial data.

4.1.3 Chapter outline

Chapter 4 starts with a review of the statistical methodology and parenting trials literature that informed my approach for incorporating IVs into the MI-BT procedure for mediation analysis. Following this, in Section 4.3, I propose a new approach, called the IV-MI-BT approach, to estimate causal direct and indirect treatment effects under both measured and unmeasured confounding in the present of missing values and taking account of trial design features. Section 4.4 applies the IV-MI-BT approach to re-analyse the SPOKES trial data and discusses the substantive findings. In particular, SPOKES mediation analysis findings from Chapter 3 (assuming no unmeasured confounding) and Chapter 4 (allowing for unmeasured confounding) are contrasted and differences are interpreted. Chapter 4 concludes with a discussion of the benefits and limitations of the IV-MI-BT approach.

4.2 Review of related statistical methodology

This section reviews the statistical methodology and literature of parenting trials, which informed my development of the IV-MI-BT approach. Specifically, I will describe the construction of an instrumental variables estimator under a linear model and summarize its statistical properties. This is followed by a description of the statistical assessment of treatment effect moderation by baseline variables in trials and a review of parenting trials literature to identify a list of candidate IY Parenting Programme moderators. Finally, I will review Multiple Imputation approaches for analysis models that contain interaction terms to ensure that a proper imputation procedure will be applied for handling missing data.

4.2.1 The two-stage least squares estimator (2SLS)

To motivate the application of the IV estimation and achieve a good understanding of the IV approach, this section demonstrates the methodological foundation of the IV estimator to deal with the endogeneity bias in the linear regression model using an example of a single IV for one endogenous explanatory variable, followed by an introduction to the methodological details of constructing a 2SLS estimator and a summary of its statistical properties.

4.2.1.1 Understanding instrumental variables regression

The standard IV regression model is obtained by augmenting the standard linear regression model with a model for the endogenous explanatory variable and the instrumental variable (Wooldridge, 2002), considering the simple one endogenous explanatory variable and one instrument scenario, as follows:

$$Y = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots \delta_r X_r + \beta T + \varepsilon_Y \qquad \text{Equation 4-1}$$

$$T = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_r X_r + \zeta_1 \mathrm{Z}_1 + \varepsilon_T \qquad \text{Equation 4-2}$$

where Y is the outcome, the matrix $\mathbf{X}$ includes the set of r exogenous variables (regressors) and the unity vector, T is a single endogenous variable, Z_1 is an instrumental variable (excluded regressor). δ_0 and θ_0 are intercepts in respective models. $\delta_1, \dots, \delta_r$ are the regression coefficients of the included exogenous variables and β is the regression coefficient of the endogenous variable in the model of Y. $\theta_1, \dots, \theta_r$ are the regression coefficients of the included exogenous variables and ζ_1 the regression coefficient of the instrumental variable in the model of T. ε_Y and ε_T are respective error terms with

expectation zero. Importantly, ε_Y is correlated with T, $cov(T, \varepsilon_Y) \neq 0$ but it is uncorrelated with the instrument Z_1 and each exogenous X_j, where j indexes $X_1, X_2, \dots, X_r$.

It should be mentioned that there are no restrictions on the distributions of Z_1 and X_j. Z_1 and X_j can be continuous, discrete or binary. An exogenous variable serves as its own instrumental variable, but conventionally we just refer to the instrument for the endogenous explanatory variable. More specifically, if Z_1 satisfies two conditions (IV definition):

IV1. Z_1 is uncorrelated with ε_Y, i.e. $cov(Z_1, \varepsilon_Y) = 0$; and
IV2. Z_1 is correlated with T, i.e. $\zeta_1 \neq 0$,

Then, we simply call Z_1 an instrument for T. Since the X_j are uncorrelated with ε_Y, i.e. $cov(X_j, \varepsilon_Y) = 0$ for $j = 1, 2, \dots, r$, they serve as their own instrumental variables.

As described in Section 4.1.2, an estimate of the coefficient β obtained via the OLS method based on the variables contained in Equation 4-1 alone is biased as $cov(T, \varepsilon_Y) \neq 0$. The IV assumptions (IV1 and IV2) solve the identification problem for β in Equation 4-1 (Identification of a parameter from observed data follows once we can express β in terms of population moments in observable variables). The proof of the estimator's identifiability is straightforward:

Let us write Equation 4-1 in a matrix format

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\lambda} + \varepsilon_Y \qquad \text{Equation 4-3}$$

where $\mathbf{Y}$ is outcome, the constant, exogenous variables and the endogenous variable are included into $\mathbf{X}$ so that it is a $1\times(r+2)$ vector $\mathbf{X} = (1, X_1, \dots, X_r, T)$, $\boldsymbol{\lambda}$ is a $(r+2)\times 1$ vector $(\delta_0, \delta_1, \dots, \delta_r, \beta)^{\mathrm{T}}$ ($^{\mathsf{T}}$ superscript indicates transposition) and ε_Y is an error term.

I write the $1 \times (r+2)$ vector of all exogenous variables as $\mathbf{Z} = (1, X_1, \dots, X_r, Z_1)$. The linear model assumptions that $\mathrm{E}(\varepsilon_Y) = 0$, $cov(X_j, \varepsilon_Y) = 0$ for $j = 1, 2, \dots, r$, and the IV1 assumption $cov(Z_1, \varepsilon_Y) = 0$ imply that

$$\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\varepsilon_Y) = 0 \qquad \text{Equation 4-4}$$

Multiplying Equation 4-3 through by $\mathbf{Z}^{\mathrm{T}}$, and taking expectations, it gives

$$[\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})]\boldsymbol{\lambda} = \mathrm{E}(\mathbf{Z}^{\mathrm{T}}Y) \qquad \text{Equation 4-5}$$

where $\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})$ is $(r+2) \times (r+2)$ and $\mathrm{E}(\mathbf{Z}^{\mathrm{T}}Y)$ a $1 \times (r+2)$ vector. Equation 4-5 represents a system of $(r+2)$ linear equations with $(r+2)$ unknowns $\delta_0, \delta_1, \ldots, \delta_r, \beta$. This system has a unique solution if and only if the $(r+2) \times (r+2)$ matrix $\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})$ has full rank; that is,

$$\operatorname{rank} \mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X}) = r + 2 \qquad \text{Equation 4-6}$$

Equation 4-6 rules out perfect collinearity in Z and it holds only if $\zeta_1 \neq 0$. In this case the solution is

$$\boldsymbol{\lambda} = [\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})]^{-1}\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{Y}) \qquad \text{Equation 4-7}$$

The expectations $E(\mathbf{Z}^{\mathrm{T}}\mathbf{X})$ and $E(\mathbf{Z}^{\mathrm{T}}\mathbf{Y})$ can be consistently estimated using a random sample of variables $(\mathbf{X}, Y, Z_1)$, and so Equation 4-2 identifies the vector $\boldsymbol{\lambda}$. Given a random sample (size N) from the population, an *instrumental variable estimator* (Hansen, 1982) of $\boldsymbol{\lambda}$ is

$$\hat{\boldsymbol{\lambda}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{Y} \qquad \text{Equation 4-8}$$

where $\mathbf{Z}$ and $\mathbf{X}$ are $N \times (r+2)$ data matrices and $\mathbf{Y}$ is the $N \times 1$ data vector. The consistency of the estimator is immediate from Equation 4-7 and the law of large numbers.

4.2.1.2 The method of two stage least squares (2SLS)

Consider that there are multiple IVs, $Z_1, Z_2, \ldots, Z_k$ for T, which indicates that each Z_i is uncorrelated with ε_Y and each Z_i has some partial correlation with T for $i = 1, 2, \ldots, k$. In fact, any linear combination of $X_1, \ldots, X_r, Z_1, \ldots, Z_k$ is uncorrelated with ε_Y and can serve as an instrument for T. Now, define all the exogenous variables as $\mathbf{Z} = (1, X_1, \ldots, X_r, Z_1, \ldots, Z_k)$ that is a $1 \times (k + r + 1)$ vector. Out of all possible linear combinations of $\mathbf{Z}$ that can be used as an instrument for T, the method of 2SLS chooses the one that is most highly correlated with T, in which case, it is given by the linear regression of T on $\mathbf{Z}$.

$$T = \theta_0 + \theta_1 X_1 + \cdots + \theta_r X_r + \zeta_1 \mathrm{Z}_1 + \cdots + \zeta_k \mathrm{Z}_k + \varepsilon_T \qquad \text{Equation 4-9}$$

Under the standard assumption that there are no exact linear dependencies among the exogenous variables, OLS provides a consistent estimate of parameters in Equation 4-9. We can get the OLS fitted values:

$$\hat{T} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \cdots + \hat{\theta}_r X_r + \hat{\zeta}_1 Z_1 + \cdots + \hat{\zeta}_k Z_k \qquad \text{Equation 4-10}$$

Define $\widehat{\mathbf{X}} = (1, X_1, \ldots, X_r, \hat{T})$. Using $\widehat{\mathbf{X}}$ as the instruments for $\mathbf{X}$, it gives the IV estimator of $(r+2) \times 1$ vector $\boldsymbol{\lambda}$.

$$\hat{\boldsymbol{\lambda}} = (\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{X})^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}Y \qquad \text{Equation 4-11}$$

Note that $\widehat{\mathbf{X}}$ can be expressed as $\widehat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{X} = \mathbf{P_z}\mathbf{X}$ from OLS regression Equation 4-2, where $\mathbf{P_z} = \mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathrm{T}}$ is the projection matrix. Therefore, $\widehat{\mathbf{X}}^{\mathrm{T}}\mathbf{X} = \mathbf{X}^{\mathrm{T}}\mathbf{P_z}\mathbf{X} = (\mathbf{P_z}\mathbf{X})^{\mathrm{T}}(\mathbf{P_z}\mathbf{X}) = \widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}}$. Plugging this expression into Equation 4-11 gives the 2SLS estimator

$$\hat{\boldsymbol{\lambda}} = (\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}Y \qquad \text{Equation 4-12}$$

The 2SLS estimator $\hat{\beta}$ for T can be constructed by fitting two regressions:

The first stage OLS regression: Regress T on all exogenous variables $Z_1, Z_2, \ldots, Z_k$ and $X_1, X_2, \ldots, X_r$ to generate the fitted value $\hat{T}$ as in Equation 4-10.

The second stage OLS regression: Regress Y on the included exogenous variables $X_1, X_2, \ldots, X_r$ and the fitted value $\hat{T}$, then we get the 2SLS estimator $\hat{\beta}$ for T.

In fact, the 2SLS estimator in Equation 4-12 and the IV estimator in Equation 4-8 are identical when there is only one IV for T. The 2SLS procedure is easily implemented using IV routines in commercial software packages such as Stata (StataCorp, 2009).

4.2.1.3 Statistical properties of the 2SLS estimator

The 2SLS estimator is an asymptotically unbiased (*consistent*) estimator of the relationship between the endogenous explanatory variable and the outcome. This means that the 2SLS estimator is close to the causal effect of interest when the sample size (N) increases. Considering the population model for outcome in Equation 4-1 and the model for the endogenous variable in Equation 4-9, where $\mathbf{X}$ is a $1\times (r+2)$ vector as $\mathbf{X} = (1, X_1, \ldots, X_r, T)$

and $\mathbf{Z}$ is a $1 \times (k + r + 1)$ vector as $\mathbf{Z} = (1, X_1, \dots, X_r, Z_1, \dots, Z_k)$. The consistency property of the 2SLS estimator requires two assumptions: Assumption (IV1)**:** $\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\varepsilon_Y) = 0$. It implies that each Z_i is uncorrelated with ε_Y for $i = 1, 2, \dots, k$ and the expectation of ε_Y is zero. Assumption (IV2): The *rank condition*: $rank\ \mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X}) = r + 2$. More specifically, it means $\mathbf{Z}$ is sufficiently linearly related to $\mathbf{X}$ so that $\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})$ has full rank. Equivalently, it means that at least one ζ_i for $i = 1, 2, \dots, k$ in Equation 4-9 is nonzero. To achieve this, a necessary condition is called the *order condition* ($k > r$), which means that the number of IVs should be at least equal to the number of endogenous variables.

Furthermore, the 2SLS estimator is asymptotically normally distributed under assumptions IV1 and IV2, and an additional homoscedasticity assumption $\mathrm{E}(\varepsilon_Y^2|\boldsymbol{Z}) = \sigma^2$ (**Assumption 3**). Asymptotic normality of $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ simply follows from the central limit theorem under Assumption IV1 and a mild finite second-moment condition, which is usually a realistic thing to hope for (Feller, 2008). Assumption 3 is only needed to determine the variance. Under assumptions (IV1), (IV2) and (3), $\sqrt{N}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normally distributed with mean zero and asymptotic variance matrix $\sigma^2\{\mathrm{E}(\mathbf{X}^{\mathrm{T}}\mathbf{Z})[\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{Z})]^{-1}\mathrm{E}(\mathbf{Z}^{\mathrm{T}}\mathbf{X})\}^{-1}$ (Wooldridge, 2002). As a crucial point, while this project relies on the consistency property of the IV estimator when generating the point estimate of the mediation effects of interest, I am not exploiting the asymptotic variance of the 2SLS estimator, since I later expand the linear model to include random effects.

A potential pitfall of the IV 2SLS is that the 2SLS standard errors have a tendency to be "large". We may end up in a situation where the 2SLS standard errors are much larger than the OLS standard errors and nothing is significant. I now show how this *variance inflation* arises: As mentioned above, the 2SLS estimator is asymptotically normal with asymptotic variance under certain assumptions. Focusing on the asymptotic variance (variance for large N) of the 2SLS estimator $\hat{\beta}$, we have

$$\mathrm{var}(\hat{\beta}) \approx \frac{\sigma^2}{\widehat{SSR}} \qquad \text{Equation 4-13}$$

where $\widehat{SSR}$ is the sum of squared residuals from the regression: $\hat{T}$ regressed on $(1, \hat{X}_1, \dots, \hat{X}_r)$.

If X are exogenous variables, then $(1, \hat{X}_1, \ldots, \hat{X}_r)$ = $(1, X_1, \ldots, X_r)$. From the definition of R-squared, we can write

$$\widehat{SSR} = \widehat{SST}(1 - \hat{R}^2) \qquad \text{Equation 4-14}$$

where $\widehat{SST}$ is the total sum of squares of $\hat{T}$ is the sample, $\widehat{SST} = \sum_{i=1}^{N}(\hat{T}_i - \bar{\hat{T}})$ and $\hat{R}^2$ is the R-squared from regression: $\hat{T}$ regressed on exogenous variables $(1, X_1, \ldots, X_r)$. $\hat{T}$ is the fitted value of regression Equation 4-9.

Equation 4-14 tells us that the (asymptotic) variance of the estimator consists of two parts: the $\widehat{SST}$ that measures the total variation in $\hat{T}$ and the $(1 - \hat{R}^2)$ viewed as a measure of multiplicity. Actually, the multiplicity is the source of the OLS standard error, so that the additional source of large variance of the 2SLS estimator is the small $\widehat{SST}$. If sample size increases, $\widehat{SST}$ gets larger. Therefore, a common approach to improving the precision of the 2SLS estimator is increasing the sample size. Moreover, in the 2SLS approach, if T is only weakly related to the IVs, then the explained sum of squares from regression Equation 4-9 can be quite small, causing a large asymptotic variance for $\hat{\beta}$. If T is highly correlated with the predictors in Equation 4-9, then $\widehat{SST}$ can be almost as large as SST (total sum of squares of T), and this fact reduces the 2SLS variance estimate. $\widehat{SST}$ is the same as the explained sum of squares from Equation 4-9. Thus a larger first stage model R-square implies a smaller variance inflation of the 2SLS estimator.

If the explanatory power of instruments is poor, then $\hat{R}^2$ is close to one. This means when the excluded instruments (IVs) have little explanatory value for T once $X_1, \ldots, X_r$ have been controlled for, in which case the $1 - \hat{R}^2$ can be very small and may lead to a large asymptotic variance of the 2SLS estimator. The instrument relevance (the correlation between instruments and endogenous explanatory variables) can be assessed by calculating the partial R-squared for instruments in Equation 4-9. For one endogenous variable T, the partial R-squared measures the proportion of T's variance explained by instrument $\boldsymbol{Z}$ which is not explained by $\boldsymbol{X}$: that is, partial R-squared $= (\mathrm{SSE}(\mathbf{X}) - \mathrm{SSE}(\mathbf{X}, \mathbf{Z}))/\mathrm{SSE}(\mathbf{X})$, where SSE means Error Sum of Squares. Shea's (Shea, 1997) "partial R-squared" is a more general form of partial R-squared that can be used to measure the instrument's relevance when multiple endogenous variables are investigated. For each endogenous variable (T_i), Shea's partial R-

squared is the squared correlation between the component of T_i that is orthogonal to the other explanatory variables and the component of the predicted values of T_i that is orthogonal to the predicted values of the other explanatory variables. Note that Shea's partial R-squared equals the general partial R-squared (Bound et al., 1995) when there is only one endogenous variable, but at least one exogenous variable, as in Equation 4-1. Shea's "partial R-squared" is commonly used as one of the diagnostic criteria to assess instrumental variables, as a larger Shea's "partial R-squared" indicates small variance inflation of the IV estimator. Four regression steps have been described in Shea's original paper to calculate partial R-squared, which has also been implemented in Stata using the written command ivreg2 (Baum et al., 2003).

Assuming that the IV assumptions are not violated, the IV estimator $\hat{\beta}$ is only asymptotically unbiased, meaning that some bias will exist when the estimator is used in smaller samples. This *finite sample bias* appears because the relationship between the instrumental variable and the endogenous variable is generally unknown and has to be estimated from the sample observations. Besides the asymptotic variance inflation, the IV 2SLS approach faces considerable challenges when the instrument is weakly correlated with the endogenous explanatory variable, and is then referred to as a 'weak instrument'. For weak instruments, the finite sample distribution of the IV estimator can depart dramatically from the asymptotic normal distribution. For example, Buse (Buse, 1992) approximated the exact finite sample distribution of IV and showed that IV is biased in the direction of OLS, with the bias increasing as instruments grow less relevant. In addition, in a survey of weak instrumental variables, Stock, Wright and Yogo (Stock et al., 2002) showed that weak instruments lead to the 2SLS estimator having a non-normal sampling distribution, regardless of sample size, so that inferences derived from large sample properties are unreliable. With moderate sample size and a weak instrument, this finite sample bias can be substantial (Sawa, 1969).

For weak instruments, the violation of IV assumptions has a large impact. If variable Z has some correlation with the error term ε_Y, then Z violates assumption IV1 and is not a valid IV. As a result, the IV estimator is inconsistent. The weaker the (false) IV variable Z, the worse the inconsistency. To see this, consider the simple IV model in Equation 4-1 and Equation 4-2: the probability limit (plim) of the IV estimator can be written as

$$\operatorname{plim} \hat{\beta} = \beta + \left(\frac{\sigma_{\varepsilon_Y}}{\sigma_T}\right)[\operatorname{Corr}(Z, \varepsilon_Y)/Corr(Z, T)] \quad \text{Equation 4-15}$$

where $\operatorname{Corr}(.\ ,\ .)$ denotes correlation. From this equation, we see that if Z and ε_Y are correlated, the inconsistency in the IV estimator gets arbitrarily large as $\operatorname{Corr}(Z, T)$ gets close to zero. Thus small correlations between Z and ε_Y can cause severe inconsistency and therefore severe finite sample bias if Z is only weakly correlated with T. This point is of practical importance, as investigators often cannot rule out minor violations of IV1.

Information on the finite sample bias is contained in the F-statistics of the regression of endogenous variables on instrumental variables and other regressors as in Equation 4-9 (first stage F-statistics). An F-value not far from 1 indicates a large finite sample bias, whereas a value of 10 seems to be sufficient for the bias to be negligible (Staiger and Stock, 1997). Stock and Yogo (Stock and Yogo, 2002) offered two alternative quantitative definitions of weak instruments. The first definition is that a group of instruments is weak if the bias of the IV estimator, relative to the bias of the OLS estimator, could exceed a certain threshold *b*, for example relative bias of 10%. The second is that the instruments are weak if the conventional α-level Wald test based on IV statistics has an actual size that could exceed a certain threshold *r*, for example *r* = 10% when α = 5%. Each of these definitions yields a set of population parameters that define weak instruments: that is, a "weak instrument set." They have also tabulated the critical values to enable the use of the Cragg-Donald F-statistic (Cragg and Donald, 1993) when there are multiple endogenous regressors, to test whether given instruments are weak. Based on Stock and Yogo's research, the classical criterion of an F-value greater than 10 leads to roughly 10% relative bias. Sawa (Sawa, 1969) found that the relative bias from the 2SLS method is asymptotically approximately equal to $1/F$, which provides a more straightforward method for quantifying the bias of the IV 2SLS estimator in practice.

To sum up, compared with the OLS estimator, the 2SLS estimator tends to have larger standard errors, where the variance inflation can result from low relevance between the endogenous variable and the instruments. The 2SLS estimator is a consistent estimator that relies on large sample size, and its finite sample bias is exacerbated by weak instruments. Tools (R-squared and F-statistics) have been suggested for assessing the strength of

instrumental variables: Shea's partial R-squared can be used as an indicator of the precision of the 2SLS estimator (the size of asymptotic variance); the F-statistics are applied to diagnose weak instruments for the purpose of monitoring finite sample bias. An ideal instrument needs to satisfy the assumptions that instrument (Z) has no correlation with error ε_Y (assumption IV1) and is strongly related to the endogenous variables. As a requirement of both precision improvement and bias reduction, large sample size is crucial for IV 2SLS estimation.

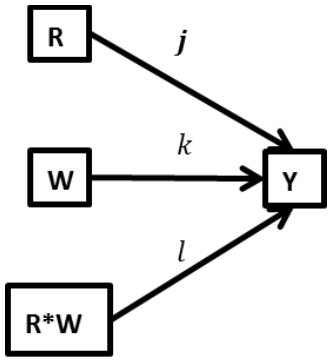
4.2.2 Treatment effect moderation in parenting trials

To address the unmeasured confounding between the post-randomisation mediator (endogenous variable) and the outcome, this project will develop IV methodology for causal mediation analysis of RCTs. Baseline covariates that modify the effect of randomised treatment on the mediator (i.e. covariate by treatment interactions) provide us with valid instruments for this purpose, assuming that the respective treatment by covariate interactions has no direct effect on the outcome. In this section, I will review the statistical methods for assessing treatment effect modification by baseline variables (moderation) and summarise the results of previous moderation analyses of trials of parenting intervention. This will provide a list of putative instrumental variables to be used in later causal mediation analyses of parenting trials.

4.2.2.1 Basic statistical methodology for moderation analysis

The moderator analyses in trials are concerned with identifying subgroups of patients who are responding differentially to the intervention, and thus they may lead to a better understanding of the causal processes operating in these subgroups (Hinshaw, 2002). Statistically a moderation effect is represented by an interaction between the baseline characteristic and the intervention, which allows moderators to differentially predict outcome across treatment groups. In contrast, predictors predict outcome regardless of treatment condition. Linear regression or ANCOVA/ANOVA methods are typically used for testing respective treatment by covariate interactions (Baron and Kenny, 1986). A moderator model includes three causal paths (j, k and l) as depicted in the diagram in Figure 4-4.

Figure 4-4 Moderator model



Path "j" represents the main effect of the treatment R on the outcome Y; path "k" represents the main (predictive) effect of the moderator W; and path "l" is the interaction effect of the predictor and the moderator ($R \times W$) on outcome Y. The interaction term $R \times W$ (path "l") needs to exist to support the occurrence of treatment effect moderation. It may also be of interest to assess the main effect of W (to establish a variable as a predictor) but resulting coefficients need to be interpreted with care, as their meaning depends on the particular coding employed to represent interactions (for more, see 4.2.2.3). In the context of IV mediation analyses, the focus is on the interaction effects of $R \times W$ on the intermediate outcome M, where W is a baseline covariate. For the purpose of methodological illustration, I use Y to reflect the dependent variable in the regression model following general convention.

The regression model with interaction term is written as:

$$Y = \beta_0 + \beta_1 R + \beta_2 W + \beta_3 RW + \varepsilon \qquad \text{Equation 4-16}$$

where the product term RW represents the interaction $R \times W$, β_0 is the intercept of the model, $\beta_1, \beta_2, \beta_3$ are further regression coefficients and ε is the error term. Moderation analysis is typically concerned with testing whether coefficient β_3 is zero.

In general, within a mediation model, baseline covariates can be evaluated as moderators (1) that influence the size of the direct treatment effects (treatment-outcome moderator), (2) that modify the effect of the treatment on the level of mediator (treatment-mediator moderator), (3) or that change the effect of the mediator on the final outcome (mediator-

outcome moderator) (MacKinnon, 2008). If the moderator of the treatment effect on the mediator (moderator 2) is believed to serve as an IV for the endogenous mediator, then in order to fulfil the IV assumptions, this baseline covariate should not moderate the direct effect of the treatment on the distal outcome (moderator 1; exclusion restriction for interaction IVs in mediation models). It is possible to model moderation of the mediator-outcome relationship (moderator 3) but resulting interaction terms would lead to further endogenous explanatory variables in the outcome model. I thus chose not to explicitly model variability in the causal mediator effect.

4.2.2.2 Moderation analysis conducted in studies of parenting programmes

In the field of parenting interventions for child problem behaviour, a few studies have examined moderation by baseline covariates of the total effect of treatment on child outcome (Eyberg et al., 2008, Reyno and McGrath, 2006, Reid et al., 2004, Beauchaine et al., 2005, Gardner et al., 2009, Kling et al., 2010). The studies' results provided a set of candidate moderators (children's initial levels of conduct problems, mothers' initial levels of critical parenting, maternal educational level, single parenthood, marital adjustment, maternal depression, paternal substance abuse, child comorbid anxiety/depression, and child age) that can be considered as providing interaction IVs for causal mediation analysis in the presence of unmeasured confounding of the mediator-outcome relationship:

In a large-scale effort (*n*=882), Reid et al. (Reid et al., 2004) examined possible baseline parent and child moderators of outcome in trial of the "Incredible Years" Parent Training Programme. Structural equation modelling (SEM) was used to model the effects of the training programme on child outcomes. Results suggested moderating effects for children's initial levels of conduct problems and mothers' initial levels of critical parenting. The children who had high baseline levels of conduct problems and whose mothers had high initial levels of critical parenting benefited most from the program. Beauchaine et al. (Beauchaine et al., 2005) combined data from six RCTs of the "Incredible Years" programme including 514 children aged 3 to 8.5. Latent growth curve modelling (LGM) was used to examine the predictors and moderators of treatment effects for children with oppositional defiant disorder or conduct disorder. A LGM model was used to describe individual trajectories over time and capture the information by underlying growth factors. The model was then expanded for the purposes of moderation assessment by including pre-treatment parent

and child characteristics and their interactions with treatment condition as explanatory variables in the models for the growth factors. The moderator analysis results suggested that marital adjustment, maternal depression, paternal substance abuse and child comorbid anxiety/depression each moderated treatment response. IY training interventions achieved better effects in children of low marital adjustment mothers, children of mothers with maternal depression, children of parents with substance abuse histories, and children who scored below the median on the anxiety/depression scale. Similarly, moderation hypotheses were examined with LGM in a large-sample (*n*=731) RCT of a family-centred intervention (the Family Check-Up) for problem behaviour in early childhood (2- to 3-year-olds: (Gardner et al., 2009). Moderation of the effect of treatment on the rate of change in problem behaviours was captured by a series of interaction effects between intervention status and the covariates. Maternal educational level and single parenthood were revealed as the two moderators of intervention effectiveness. Caregivers with the lowest educational levels were more responsive to the family-centred intervention, as were two-parent families. Kling *et al.* tested nine family characteristics as putative moderators using multiple regression analysis (Kling et al., 2010) in a parent management training (PMT) RCT including parents of 159 children (aged 3 to 10) with conduct problems. Results suggested that children of younger mothers benefit the most from the intervention.

4.2.2.3 The method of orthogonalisation for interaction terms

The moderation analysis introduced in Section 4.2.2.1 may encounter several problems, including *collinearity* due to high correlation between product and main effect terms and uninterpretable regression parameter estimates. Collinearity means that within the set of explanatory variables, one or more of the variables are highly predicted by one or more of the other explanatory variables. The consequence of collinearity is unstable regression estimates, which indicates that minor fluctuations in the sample will have a major impact on the estimates. Under orthogonal conditions, when the interaction term is entered into a model, the partial regression coefficients representing the magnitudes, directions, and significances of the main effect variables remain precisely the same as they were before the interaction was included. An orthogonalisation approach proposed by (Lance, 1988) has the above merits. It is essentially an OLS procedure in which a product term is regressed onto its respective first-order effect(s).

$$RW = \theta_1 R + \theta_2 W + \varepsilon_{RW} \quad \text{Equation 4-17}$$

The residuals (ε_{RW}) of this regression are then used to represent the interaction.

$$\varepsilon_{RW} = RW - \widehat{RW} = RW - \widehat{\theta_1} R - \widehat{\theta_2} W \quad \text{Equation 4-18}$$

The final regression model is now

$$Y = \beta_0 + \beta_1 R + \beta_2 W + \beta_3 \varepsilon_{RW} + \varepsilon_Y \quad \text{Equation 4-19}$$

The variance of this new orthogonalised interaction term contains the unique variance that fully represents the interaction effect and is independent of the first-order effect variance (as well as general error or unreliability). The orthogonalising has a number of inherent advantages for regression analyses. First, estimates of regression coefficients for orthogonalised product terms are numerically stable. Second, the interpretation of the regression coefficients of the first-order effect terms remains unchanged when the interaction term is entered. Third, since treatment allocation is at random, such residual centering ensures full independence between the product term and its constituent main effects as well as between *R* and baseline variables.

In the context of IV mediation analysis, using $R \times M$ interaction term as IV without orthogonalisation will yield an estimation of the effect of treatment on the mediator for a certain level of the baseline variable. In contrast, orthogonalising the IV (interaction term) will preserve the estimate of the average treatment effect on the mediator (α). In other words, using orthogonalised IVs in the mediation analysis will address unmeasured confounding without changing the meaning of the causal parameter estimates. Thus, this orthogonalisation procedure will be applied to raw candidates' IVs (interaction terms) before inclusion in mediation models.

4.2.3 Methodology of Multiple Imputation in instrumental variables setting

In this Chapter, I propose to embed IV estimation into the MI-BT approach introduced in Chapter 3. Therein, the Multiple Imputation approach was described to handle missing data in mediation investigations with observed confounding. Whilst there has been considerable

research into methods of imputation, I am not aware of specific research into appropriate Multiple Imputation models for IV estimation. Although further research is required into Multiple Imputation methods to address missing data issues in IV estimation (Palmer et al., 2012), the well-known rules for generating proper MI will help us to construct imputations for mediation analysis models with IVs. The two rules are: (1) congeniality between imputation and analysis models; (2) variables included in the analysis model must be included in the imputation model. Following the latter rule, the IVs must be included in the imputation model. Application of the first rule in combination with the IV exclusion restriction assumption suggests that we use IVs to predict missing values in the target explanatory variables (the mediators), but exclude such IVs from predicting missing values in the outcome variables.

As described in Section 4.2.2, the proposed IVs are treatment by baseline covariate interaction terms, while the baseline covariate itself (the first order variable) is also included in the mediation model as an observed confounder. In the following paragraphs, I will review relevant MI techniques for interaction terms in the analysis model (White et al., 2011) and recommend appropriate techniques for use in IV mediation analysis in this project:

The simplest approach is *passive imputation,* which only imputes the first order variables and generates predictions for missing interaction terms by taking respective products. For example, missing values are present in child's gender (W) and the randomisation by child gender interaction ($R \times W$) is included in the analysis model of positive parenting outcome (M). The passive approach imputes the missing data in child gender in the usual way and calculates the treatment-gender interaction passively - that is, by multiplying the imputed child's gender values with the treatment variable values. Since under this approach variables from the analysis model (interaction terms) are excluded from the imputation model, rule (2) is not adhered to and the cost is bias of relevant terms in the analysis model and a loss of power to detect interactions. In this project, I am interested in detecting moderation effects on the mediator, and would not wish for such assessments to be biased towards zero.

Method has been developed to improve the passive approach. Considering the same positive parenting outcome (M) with a treatment by child's gender interaction ($R \times W$) example, the presence of the interaction term ($R \times W$) in the analysis model means that the

association between positive parenting outcome (M) and child's gender (W) may differ between treatment groups (R). Thus, the interaction between positive parenting outcome and treatment groups ($R \times M$) should be included in the imputation models of positive parenting outcome (M) and child's gender (W). This new approach is called the *improved passive approach*. The improved passive approach relies on correct specification of the imputation models, so that the interaction terms to be included in the analysis model and imputation model need to be decided cautiously. As the number of variables increases, it becomes harder to find and estimate correct passive imputation models.

It is often hard to specify the imputation model, as its true form is non-standard. Instead of aiming to find the true imputation model, an alternative approach relies on finding an imputation model that is 'congenial' to the analysis model but not necessarily correctly specified (Rule 1). A popular choice of the larger model is the multivariate normal distribution. Even though the multivariate normal may be mis-specified when some variables are categorical, Schafer presents evidence that procedures based on a multivariate normal assumption perform well under this sort of model mis-specification (Schafer, 1997). The "just another variable" (JAV) approach, as its name suggests, imputes the interaction term as another variable. This approach is based on a multivariate normal model for the first order variables (R, M, W) jointly with the interaction term ($R \times W$), which requires that each variable is imputed using a linear regression.

A simpler congenial imputation approach involves separate imputation within each treatment group to allow the relationship among all the variables to differ between treatment groups. However, given the relatively small sample size of SPOKES (n=112) and the large amount of variables in the imputation model, separating imputation in treatment groups is too ambitious and the imputation model may encounter convergence failure.

In summary, I have illustrated the passive, improved passive, JAV and separate imputation in each group as approaches for imputing covariates whose interactions are included in the analysis model. None of them is flawless. In the current MI for IV (interaction term) mediation analysis, the interaction terms (IVs) will be orthogonalised (see 4.2.2.3). The question is in fact how to perform MI for analysis models that contains orthogonalised interaction terms. The orthogonalised term renders the passive approach and the improved

passive approach inappropriate due to its independence to the first order variable. As reviewed above, the separate imputation in each group is practically infeasible due to the small sample size and the large amount of variables in the imputation model. Without orthogonalisation, the JAV approach may not be able to ensure the product nature of the interaction term. However, it is no longer an issue when the interaction term is orthogonalised, via which the associations between the interaction term and the first order variables are removed. Therefore, in this project, the JAV approach is selected for imputing variables whose orthogonalised interaction terms are included in the IV causal mediation analysis model.

4.3 The new IV-MI-BT combined approach for mediation analysis

To relax the assumption of no unmeasured confounding of the mediator-outcome relationship required by the MI-BT approach proposed in the previous chapter, I developed the IV-MI-BT approach by incorporating the 2SLS IV estimation into the MI-BT approach in this section. The new approach enables mediation analysis to allow both observed and unobserved confounders, account for the hierarchical structures implied by parenting programme trials and in the presence of missing data. This section mainly focuses on how to construct a two-stage maximum likelihood type IV estimator (MI-2SML estimator) of causal mediation parameters and the steps involved in the implementation of the combined IV-MI-BT approach.

4.3.1 A two-stage ML type IV estimator of the causal mediation parameter

4.3.1.1 IV Mediation model and two-stage maximum likelihood type estimator

IV mediation analyses allows for the existence of unobserved confounding of the mediator–outcome relationship. Assuming that there are valid IVs for the endogenous mediator in an RCT, Figure 4-3 shows the IV mediation model with both measured and unmeasured confounders of the mediator–outcome relationship.

As introduced in Chapter 3, the MI-BT approach accounts for the hierarchical data structure implied by the trial design by including random effects. Building upon the MI-BT approach of Chapter 3, respective mixed effect models are also employed for the IV mediation analysis in this chapter. Again, taking SPOKES as an example that has a three-level hierarchical data structure, the IV mediation mixed-effects linear models can be written as,

$$Y = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots \delta_r X_r + \gamma R + \beta M + \mathbf{\Psi}^{(3)} u^{(3)} + \mathbf{\Psi}^{(2)} u^{(2)} + \varepsilon_Y \quad \text{Equation 4-20}$$

$$M = \theta_0 + \theta_1 X_1 + \cdots + \theta_r X_r + \alpha R + \zeta_1 Z_1 + \cdots + \zeta_k Z_k + \mathbf{\Psi}^{(3)} w^{(3)} + \mathbf{\Psi}^{(2)} w^{(2)} + \varepsilon_M \quad \text{Equation 4-21}$$

where, Y is outcome, $X_1, X_2, \ldots, X_r$ is a set of exogenous measured baseline confounding variables, R is randomised treatment (exogenous variable), M is a single endogenous mediator, $Z_1, Z_2, \ldots, Z_k$ constitute a set of k instrumental variables for M. δ_0 and θ_0 are the respective model intercepts. γ and α are the (direct) effects of randomised treatment in the outcome model and the mediator model respectively. β is the causal effect of the endogenous variable M and $\delta_1, \delta_2, \ldots, \delta_r$ are the regression coefficients of the included baseline confounders in the outcome model. $\zeta_1, \zeta_2, \ldots, \zeta_k$ are the regression coefficients of the instrumental variables and $\theta_1, \ldots, \theta_r$ are the effects of the included baseline confounders in the mediator model. $u^{(3)}$ and $w^{(3)}$ are the level-3 random effects for child outcome and mediator respectively, and $u^{(2)}$ and $w^{(2)}$ are the level-2 random effects in the IY arm for child outcome and mediator respectively. The expectations of $u^{(3)}$, $w^{(3)}$, $u^{(2)}$ and $w^{(2)}$ are zero. $\mathbf{\Psi}^{(3)}$ is the design matrix for the level-3 random effect and $\mathbf{\Psi}^{(2)}$ is the design matrix for the level-2 random effect. ε_Y and ε_M are residuals with expectation equal to zero. ε_Y is uncorrelated with R and each X_j, but $Cov(M, \varepsilon_Y) \neq 0$, and ε_M is uncorrelated with each Z_i and each X_j, where i indexes $Z_1, Z_2, \ldots, Z_k$ and j indexes $X_1, X_2, \ldots, X_r$.

The IV assumptions are:

IV1. There are no unmeasured common causes of IVs and outcome, and no direct paths from IVs to outcome, i.e. $cov(Z_i, \varepsilon_Y) = 0$ where i indexes $Z_1, Z_2, \ldots, Z_k$.

IV2. IVs are associated with the endogenous mediator M ($\zeta_i \neq 0$, where i indexes ζ_1, ζ_2, …, ζ_k)

We also assume that the effect of the mediating variable is the same for all subjects (mediator effect homogeneity).

The maximum likelihood approach was applied to generate estimates of mixed effects models' parameters in Chapter 3. Estimation of the model parameters based on the extended mixed models that include the instrumental variables can be achieved by turning

the two-stage least square (2SLS) approach for regression models into a two-stage maximum likelihood (2SML) approach for linear mixed models. I will now describe how to construct an IV ML-type estimator following a two-stage approach:

As a result of the unmeasured confounding between mediator M and outcome Y, i.e. $cov(M, \varepsilon_Y) \neq 0$, the ML-type estimator of β_{ML} is biased in Equation 4-20. Let $\mathbf{Z} = (1, X_1, X_2, \dots, X_r, R, Z_1, Z_2, \dots, Z_k)$ and $\boldsymbol{\zeta} = (\theta_0, \theta_1, \dots, \theta_r, \gamma, \zeta_1, \zeta_2, \dots, \zeta_k)$. As defined in Equation 4-21, $cov(Z, \varepsilon_M) = 0$. The ML-type estimator of $\boldsymbol{\zeta}_{ML}$ is asymptotically unbiased for large samples if the residuals follow symmetric distribution (Section 3.3.1 of Chapter 3) and so is the projection for $E(M|\boldsymbol{Z})$. It follows from Equation 4-20 that asymptotically $E(Y|\boldsymbol{Z}) = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \cdots \delta_r X_r + \gamma R + \beta E(M|\boldsymbol{Z})$ based on the IV assumptions and the zero expectation assumptions of random effects (including residuals) in the mixed-effects models. Thus the IV ML-type of estimator $\hat{\beta}_{ML}^{IV}$ is asymptotically unbiased.

The two-stage IV ML procedure can then be applied to estimate the coefficient of the endogenous mediator M:

First stage ML regression: Fitting the multilevel mixed linear regression model in Equation 4-21, and obtaining the fitted $\widehat{M}_{ML}$, the estimated $E(M|\mathbf{Z})$:

$$\widehat{M}_{ML} = \hat{\theta}_0 + \hat{\theta}_1 X_1 + \cdots + \hat{\theta}_r X_r + \hat{\alpha} R + \hat{\zeta}_1 Z_1 + \cdots + \hat{\zeta}_k Z_k \qquad \text{Equation 4-22}$$

Second Stage ML regression: Replacing the endogenous variable M with the fitted value $\widehat{M}_{ML}$ in Equation 4-20, then fitting this modified mixed model to get the IV ML-type estimator $\hat{\beta}_{ML}^{IV}$ that estimates the average effect of mediator M on outcome Y conditional on the observed confounding variables $X = (X_1, X_2, \dots, X_r)$.

After obtaining the asymptotic unbiased estimate $\hat{\beta}_{ML}^{IV}$ of the effect of the mediator M on the outcome Y (EMO) using the two-stage ML approach, the causal mediation effects can be estimated as follows: The causal effect of the treatment on the mediator (ETM) is estimated by $\hat{\alpha}_{ML}^{IV}$ and the direct effect of treatment on outcome (DE) is estimated by $\hat{\gamma}_{ML}^{IV}$ from Equation 4-21 and Equation 4-22 respectively. The indirect effect (IE) is now estimated by the product $\hat{\alpha}_{ML}^{IV}\ \hat{\beta}_{ML}^{IV}$; and the total effect (TE) by $\hat{\gamma}_{ML}^{IV} + \hat{\alpha}_{ML}^{IV}\ \hat{\beta}_{ML}^{IV}$. Of note, the $\hat{\alpha}_{ML}^{IV}$ will preserve the same meaning and value as the corresponding estimator $\hat{\alpha}_{ML}$ that was

proposed in Section 3.3.1 of Chapter 3 providing the IVs are orthogonalised (see Section 4.2.2.3) to the treatment group variable R.

4.3.1.2 Combining the 2SML IV estimator with MI and BT

As described in Chapter 3, the MI-BT approach can generate consistent causal mediation effect estimates and provide valid statistical inferences under the assumptions of no unmeasured confounding between mediator and outcome, the less restrictive MAR missing data generating assumption, and without making assumptions regarding the distribution of the residuals. In this section, the two-stage ML-type IV estimator (2SML) will be combined with Multiple Imputation (MI) and bootstrapping (BT) to provide the combined IV-MI-BT approach. Briefly, the structure of the MI-BT procedure as illustrated in Figure 3-2 is still the same; the only change is that the Baron and Kenny (Baron and Kenny, 1986) type analysis model will be replaced by the IV mediation model and consequently the estimator will be the 2SML estimator. The following paragraphs provide further details of the combined IV-MI-BT approach.

In line with the MI-BT approach proposed in Chapter 3, MICE will be used for imputing missing values. Since the IVs feature as explanatory variables in the new IV mediation analysis model, they must also be included in the imputation model (Rule 2 from Section 4.2.3). Additionally, the IV assumptions state that there is no direct path from the IVs to the clinical outcome Y, thus the IVs are excluded from the imputation of the outcome Y to preserve congeniality of the imputation model and the analysis model (Rule 1 from Section4.2.3). As mentioned before in the introduction section of this chapter, the potential IVs are baseline covariate by randomised treatment group interaction terms. For these variables, the JAV method reviewed in Section 4.2.3 for imputing interaction terms will be applied; whilst for the other variables with missing values, their imputation proceeds as described before (Section 3.3.1.2 of Chapter 3).

A 2SML estimator is constructed for each imputed data set, i.e. $\hat{\alpha}_i^{IV}$, $\hat{\beta}_i^{IV}$, $\hat{\alpha}_i^{IV}\hat{\beta}_i^{IV}$, $\hat{\gamma}_i^{IV}$ and $\hat{\gamma}_i^{IV} + \hat{\alpha}_i^{IV}\hat{\beta}_i^{IV}$ from imputation sample i. This is repeated over Multiple Imputations (say h times) and the final IV estimates (MI-2SML estimator) of causal mediation parameters are constructed by taking means over respective estimates, i.e. $1/h\sum_1^h \hat{\alpha}_i^{IV}$, $1/h\sum_1^h \hat{\beta}_i^{IV}$, $1/h\sum_1^h \hat{\alpha}_i^{IV}\hat{\beta}_i^{IV}$, $1/h\sum_1^h \hat{\gamma}_i^{IV}$, and $1/h\sum_1^h(\hat{\gamma}_i^{IV} + \hat{\alpha}_i^{IV}\hat{\beta}_i^{IV})$ for the ETM, EMO, IE, DE and TE

respectively. These causal mediation effect estimators are consistent provided that IV assumptions hold (Section 4.3.1.1), the missing data generating process is MAR and analyses models are correctly specified (including a linear relationship between the mediator and the clinical outcome, and the covariates and the outcomes, and the absence of an $R \times M$ interaction effect on Y).

Nonparametric inferences for causal mediation effects are generated using (cluster) bootstrapping as described in Chapter 3. The bootstrap resampling procedure mimics the trial data generating process. The methods of confidence interval and test construction remain the same when considering a new estimator (MI-2SML). Thus we can approximate the sampling distribution of the MI-2SML estimator by employing the same bootstrap methods as used in Chapter 3.

4.3.2 Choosing instrumental variables for an endogenous mediator

4.3.2.1 Criteria for evaluating instrumental variables

You may have noticed that the approach introduced above assumes that there are variables that can serve as IVs for the endogenous mediator M in the mediation analysis. However, it is well-known that finding appropriate IVs is often not an easy task. To address this challenge, I suggested a strategy for constructing a list of potential instrumental variables and set up criteria to evaluate the quality of competing candidate IVs in this section.

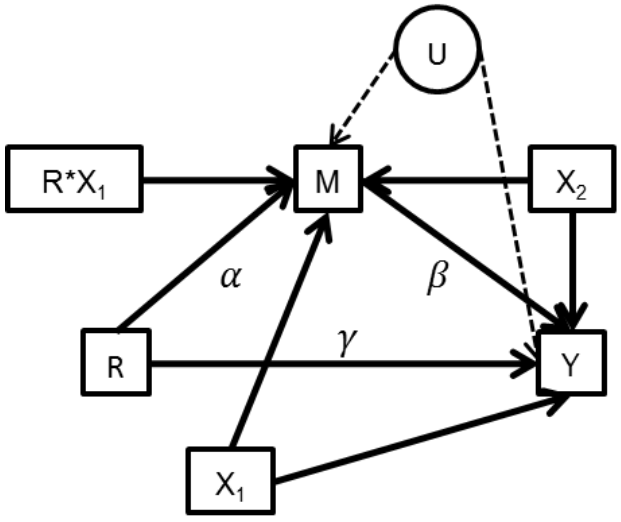
As reviewed in Section 4.1.2, the interactions between treatment randomisation and covariates (treatment effect moderators) are often used as instrumental variables (Gennetian et al., 2008, Ten Have et al., 2007, Dunn and Bentall, 2007, Dunn et al., 2013, Albert, 2008, Small, 2012, Emsley et al., 2010). Additionally, the findings of moderation analysis in studies of parenting programmes (see Section 4.2.2.2) provide the directions in which to look for treatment effect moderators. Thus, I propose that the following terms might potentially act as IVs for mediation investigations; that is, they might serve as IVs for the endogenous parenting mediator M in the linear model for child outcome Y.

a. Interactions between intervention groups and baseline parent characteristics including parental education, parental depression, and lone parent.
b. Interactions between intervention groups and baseline child characteristics including child gender, child age, and child reading ability.

c. Interactions between intervention groups and baseline parenting practice measurement including baseline expressed warmth and baseline expressed criticism.
d. Interaction between intervention groups and baseline child outcome measurement.
e. Intervention process variables such as therapy group or number of sessions attended.

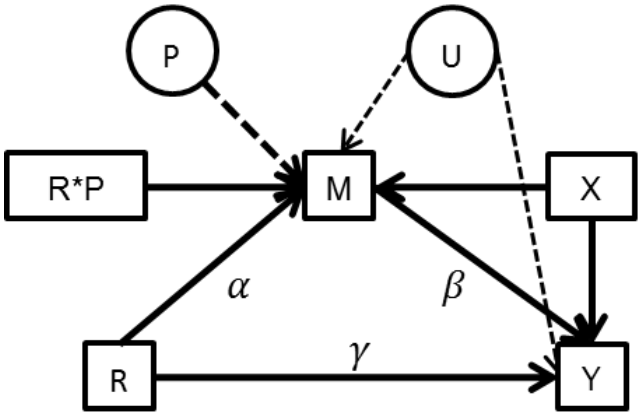
The treatment effect moderators included in items a.-d. supplying candidate interaction IVs are selected on theoretical grounds. First, interactions with randomly allocated intervention (R) will partly fulfil assumption IV1 in that the interaction term with R is also not associated with any unobserved variables. To completely fulfil assumption IV1, we only need to assume that there is no direct interaction effect on the outcome *Y*. This amounts to the assumption that the modification of the treatment effect operates solely by changing the mediator (i.e. no moderation of the direct effect of the treatment on *Y*; exclusion restriction for interaction IVs in mediation models). As shown in Figure 4-5, X_1 is the baseline variable that modifies the treatment effect on mediator M, X_2 represents baseline confounders, and interaction term $R * X_1$ is the IV. The main effect of the baseline variable (X_1) on the outcome are included in the models. Assuming that there are no direct interaction effects seems reasonable, especially when the moderator is a parent variable, as in a. and c. To fulfil assumption IV2, the endogenous parenting mediator needs to be predicted by the interaction terms; i.e. there needs to be treatment effect modification of the mediator. The latter can be assessed empirically and I have based my choice of potential interaction IVs on moderation effects suggested by previous IY Parenting Programme trials (see Section 4.2.2) and that are also measured in the project trials.

Figure 4-5 RCT single-mediation model with randomisation - baseline variable interaction IV



Intervention process variables (P) such as the therapy group of participants are only observable in the active treatment group. Their interaction with (binary) treatment R can be observed due to trial design, so that I also considered such variables as IVs (type e.). However, stronger assumptions are required for such interaction terms to serve as IVs. The same as interactions with baseline moderators, interactions with post-randomisation moderators require that the exclusion restriction holds. In addition, for IV1 to hold, they require that the latent variable "process outcome if receiving training" (a counterfactual condition in the control group) does not have a direct effect on clinical outcome Y, while the variable can affect Y via changing the parenting mediator M. In other words, we require the observed interaction term as well as the latent post-randomisation moderator to have IV properties. As shown in Figure 4-6, the process variable P does not affect outcome Y except through its effect on mediator M.

Figure 4-6 RCT single-mediation model with randomisation - intervention process variable interaction as IV



Although IV assumptions cannot be assessed empirically, there are empirical approaches that have been applied to comment on possible violations of assumption IV1 in situations where one has some prior knowledge of the direction of effects. One such "empirical check" is the comparison of the absolute size of the IV effect on the mediator with that on the outcome given the standardised effect of the mediator on the outcome is less than 1. The diagnostic follows the same logic as Glymour's paper (Glymour et al., 2012), in which it was suggested that we can falsify IV assumptions by leveraging the prior causal assumption to exclude the potential variables that should not serve as IV. Proof of this IV property diagnostic is straightforward. Let ζ be the predictive effect of the instrumental variable in question (Z) on the mediator (M), while β denotes the causal effect of mediator (M) on

outcome (Y). Based on previous research findings and the standardisation of the parameters, we might argue that $|\beta| < 1$. We allow for unobserved confounding of the causal path from M to Y. Let c be the predictive effect of Z on Y, which is not mediated by M, then $c = 0$ means Z is IV and $c \neq 0$ means Z violates IV assumptions. The diagnostic approach further assumes that the potential IV's mediated and non-mediated predictive effects, $\zeta\beta$ and c, are in the same direction (same sign). Then if $|\zeta\beta + c|$ (effect of Z on Y) $> |\zeta|$ (effect of Z on M) and $|\beta| < 1$, it follows that $|(\zeta\beta + c)|/|\beta| > |\zeta\beta + c| > |\zeta|$, which is equivalent to $|\zeta\beta| + |c| > |\zeta\beta|$ or in other words $|c| > 0$ (IV assumption violation). Therefore, if the absolute size of the IV effect on the outcome is bigger than that on the mediator, the potential IV is not valid.

As reviewed in Section 4.2, when an instrument is not strongly associated with the endogenous variable, it is referred to as a weak instrument. As mentioned before (valid but) weak instruments affect the properties of the resulting IV estimators in two respects: First, there is the risk of the IV estimator having large standard errors, which would make the estimates imprecise. Second, weak instruments can produce IV estimates that suffer from finite sample bias in the direction of the OLS estimate. In this project, the first stage F-statistics criterion will be applied for evaluating the finite sample bias and the first stage Shea's (Shea, 1997) partial R-squared will be used to monitor the variance inflation.

In summary, the following criteria were applied to choose the most practically useful IV variables from a set of potential IVs:

(1) I am willing to make respective exclusion restriction assumptions.
(2) The empirical data would be consistent with assumption IV1 if direct and indirect IV effects operated in the same direction: The IV's effect on the mediator should be no less than its effect on the outcome.
(3) The level of weak instrument bias should be acceptable: For one endogenous variable, an F-value not far from 1 indicates a large finite sample bias, whereas a value of 10 is considered to be sufficient for the bias to be negligible (roughly 10% bias relative to the OLS estimate).
(4) The variance inflation should be acceptable: There are no guidelines for determining a large enough Shea's partial R-squared. Generally, a larger R-squared value leads to less variance inflation of the IV estimator relative to that of the OLS estimator.

4.3.2.2 Comparing and combining moderators

Application of criterion (2) listed above requires comparing the size of the effect of the potential IV on the mediator with the size of its effect on the outcome. In my project, several interactions with baseline or post-randomisation moderators of treatment effect are selected as potential IV. The moderators are measured on various scales, such as binary, nominal and continuous. The size of an interaction effect between randomisation and a binary or continuous variable is straightforward to quantify. That is, to simply estimate the regression coefficient of the single orthogonalised interaction term on an outcome measure in units of its standard deviations. However, how to evaluate the size of the interaction effects between randomisation and a nominal moderator variable (more than 2 categories) is less clear and will be discussed in this section. Furthermore, the situation is more complicated for post-randomisation moderators such as therapy groups that are only observed in the treated arm, and not in the control arm. A method developed for comparing and combining interactions with post-randomisation moderators that are only observed in one randomized treatment group is demonstrated in the section.

As introduced in Section 4.2.2.1, a common approach to describing moderation in a linear regression model is

$$Y = \beta_0 + \beta_1 R + \beta_2 W + \beta_3 RW + \varepsilon \qquad \text{Equation 4-23}$$

where Y is the outcome variable, measured on a continuous scale. R is the randomised treatment group indicator and W is a moderator. The error term, ε, is assumed to have a normal distribution $\mathcal{N}(0,\sigma_\varepsilon^2)$ and to be independent of R and W. β_0, β_1 β_2 and β_3 are regression coefficients.

Kraemer (Kraemer, 2013) defines the correlation coefficient between the pairwise outcome differences and the pairwise average moderator values among randomly selected participants, one from each treatment group, as a measure of moderator effect size. More specifically, the difference of the outcome from the randomly selected anticipants pair, one from $R1$ group (e.g. treatment group) and one from $R2$ group (e.g. control group): $Y_1 - Y_2$. R is coded +1/2 for $R1$ and -1/2 for $R2$. W is standardised to have mean 0 and variance 1 (the

same standardization in both treatment groups because M and T are uncorrelated). According to the linear model

$$\Delta Y = Y_1 - Y_2 = \beta_1 + \beta_2(W_1 - W_2) + \frac{\beta_3(W_1 + W_2)}{2} + (\varepsilon_1 - \varepsilon_2) \qquad \text{Equation 4-24}$$

The correlation coefficient between ΔY and the average moderator value $W = (W_1 + W_2)/2$, $r(\Delta Y, AW)$, is suggested as a measure of moderator effect size. The effect size is invariant over linear transformations of either W or Y. It is a number between -1 and 1, with null value 0, with greater magnitudes indicating stronger moderation.

Kraemer suggested that if there is a set of moderators for the same outcome, the optimal composite moderator W^* is constructed by maximizing $r(\Delta Y, AW)$ using a multiple linear regression model, i.e. regressing ΔY on the means for the set of k moderators W_i . Each estimated regression coefficient reflects the moderation strength of the respective variable in the context of all other potential moderating variables. These regression coefficients τ_i are used as the weights for calculating the combined moderator $W^* = \sum \tau_i W_{ij}$, where i=1,2,3,…, k and j=1 for $R1$ and j=2 for $R2$.

Following the same logic, I regress ΔY on the means of the set of k therapy group dummy variables. The optimal composite moderator W^* then maximises correlation between ΔY and $\sum \tau_i AW_i$ can serve as a measure of moderation effect size for nominal moderators. These regression coefficients τ_i are used as the weights for calculating the combined moderator $W^* = \sum \tau_i W_{ij}$, where i=1,2,3,…, k and $j = 1$ for $R1$ and $j = 2$ for $R2$. The challenge in here is that W is observed only in the treated group but not in the control group (i.e. only W_1 is observed but W_2 is not observed in Equation 4-24) and AW cannot be constructed. Thus, an extension of Kraemer's moderator effect size, $r(\Delta Y, AW)$, is required.

I will now demonstrate the development of Kraemer's moderator effect size using the example of therapy groups in the treated arm. Firstly, it is necessary to provide a clear definition of the therapy group moderator effect. It is the variability in treatment effects on outcome between subgroups of people who, if being offered treatment, would participate in a certain therapy group. This therapy group moderator is counterfactual in that it cannot be observed in those who were allocated to the control group. In our case, the therapy group

was unavailable in the control arm. The moderation effect only sources from the therapy group in the treated arm. For the 10 therapy groups in the treated arm of SPOKES trial, I created 9 binary variables, taking the first therapy group as a reference group. The linear regression model with therapy groups only in the treated group can be written as

$$Y = \beta_0 + \beta_1 R + \beta_3 RW + \varepsilon \qquad \text{Equation 4-25}$$

where R is coded to 1 for the treated group and 0 for the control group. W is a binary variable for a certain therapy group. The error term, ε, is assumed to have a normal distribution $\mathcal{N}(0,\sigma_\varepsilon^2)$ and to be independent of R and W.

The difference of the outcome from the randomly selected anticipants pair, one from the $R1$ group (treated group) and one from the $R2$ group (control group): $Y_1 - Y_2$, according to the linear model, is

$$\Delta Y = Y_1 - Y_2 = \beta_1 + \beta_3 W_1 + (\varepsilon_1 - \varepsilon_2) \qquad \text{Equation 4-26}$$

Let d_1 and d_3 be the standardised regression coefficients, $d_1 = \beta_1/\sigma_\varepsilon$ and $d_3 = \beta_3/\sigma_\varepsilon$. The correlation coefficients $r(\Delta Y, AW)$ is

$$r(\Delta Y, AW) = \frac{d_3}{\sqrt{(d_3^2 + 2)}} \qquad \text{Equation 4-27}$$

To illustrate the derivation of Equation 4-27, details of the calculation are listed below:

We know that $\mathrm{corr}(\Delta Y, AW) = \mathrm{Corr}\left(\beta_1 + \beta_3 W_1 + (\varepsilon_1 - \varepsilon_2), \left(\frac{W_1 + W_2}{2}\right)\right)$.

It is also known that $\mathrm{corr}(\Delta Y, AW) = \frac{\mathrm{Cov}(\Delta Y, AW)}{\sqrt{\mathrm{var}(\Delta Y)\mathrm{Var}(AW)}}$,

where

$$\begin{aligned}\mathrm{cov}(\Delta \mathrm{Y}, \mathrm{AW}) &= \mathrm{Cov}\left(\beta_1 + \beta_3 \mathrm{W}_1 + (\varepsilon_1 - \varepsilon_2), \left(\frac{W_1 + W_2}{2}\right)\right) \\ &= \mathrm{Cov}\left(\beta_3 W_1, \left(\frac{W_1 + W_2}{2}\right)\right) \\ &= \frac{\beta_3}{2}\mathrm{Cov}\left(W_1, (W_1 + W_2)\right)\end{aligned}$$

$$= \frac{\beta_3}{2}\big(\mathrm{Cov}(W_1, W_1) + \mathrm{Cov}(W_1, W_2)\big)$$

$$= \frac{\beta_3}{2}\ \big(\mathrm{var}(W_1) + \mathrm{Cov}(W_1, W_2)\big)$$

Even though the W (therapy groups) cannot be observed in the control arm, I still assume that if participants in the control group were treated, then the probability of the participant being located to one therapy group would be the same as in the treated group, i.e. counterfactual therapy group is independent of treatment group. Following this, we get $\mathrm{Cov}(W_1, W_2) = 0$. I also standardised W_1 to mean 0 and variance 1, so that $\mathrm{var}(W_1) = 1$. Therefore,

$$\mathrm{Cov}(\Delta Y\ , AW) = \frac{\beta_3}{2}$$

$$\mathrm{corr}(\Delta Y, AW) = \frac{\beta_3}{2\sqrt{\mathrm{Var}\big(\ \beta_1\ + \beta_3 W_1 + \ (\varepsilon_1 - \varepsilon_2)\big)\mathrm{Var}\left(\frac{W_1\ + W_2}{2}\right)}}$$

$$= \frac{\beta_3}{2\sqrt{(\beta_3^2 + 2\sigma_\varepsilon^2)(\frac{1}{4})}}$$

Let d_3 be the standardised regression coefficient, $d_3 = \beta_3/\sigma_\varepsilon$. Then,

$$\mathrm{corr}(\Delta Y, AW) = \frac{d_3\sigma_\varepsilon}{\sqrt{(d_3^2\sigma_\varepsilon^2 + 2\sigma_\varepsilon^2)}} = \frac{d_3}{\sqrt{(d_3^2 + 2)}}$$

This formula only depends on the (standardised) regression coefficient of the interaction term. Although our $r(\Delta O, AW)$ is differ to Kraemer's $\mathrm{corr}(\Delta Y, AW) = \frac{d_3}{2\sqrt{(d_2^2 + \frac{1}{4}d_3^2 + 1)}}$, they hold the same characteristics that $r(\Delta O, AW)$ is invariant over linear transformations of either R or W and it is a number between -1 and +1, with null value 0, with greater magnitude indicating stronger moderation.

For the partly-observed post randomisation moderator variable, we cannot observe any values for W_2 and thus cannot construct the average AW. However, we can construct the pair difference ΔY as the outcome is observed for both the control and the treatment groups. We can estimate AW by its expectation for any subject; $\mathrm{est.}(\widehat{AW}) = 0.5 \times [W_1 + E(W \mid \text{allocated to control group})] = 0.5 \times [W_1 + E(W \mid \text{allocated to treated group})]$

under independence of R and W. This can also be written as est.$(\widehat{AW}) = 0.5 \times [W_1 +$ average$(W_2)] = 0.5 \times [W_1 +$ average$(W_1)]$, thus $\text{corr}(\Delta\text{Y}, W_1)$ is an approximation of $\text{corr}(\Delta\text{Y}, AW)$. The moderation index $\text{corr}(\Delta\text{Y}, W_1)$ could be estimated from pairs of Ys and single W_1s and moderation concept extended to include multiple (partly-observed) post-randomisation moderators. Thus, $\text{corr}(\Delta\text{Y}, \sum \tau_i W_{i1})$ is the moderation effect size for k partial-moderators (i.e. observed only in the treated group) with $i = 1, 2, \dots, k$.

In summary, the moderator effect size is $r(\Delta Y, AW)$ for singleton moderators, the best combined moderator effect size is $\text{corr}\left(\Delta\text{Y}, \sum \tau_p AW_p\right)$ for fully observed moderators and the best combined moderator effect size is $\text{corr}\left(\Delta\text{Y}, \sum \tau_q W_{q1}\right)$ for partly observed moderators with $p = 1, 2, \dots, m$ and $q = 1, 2, \dots, n$. Following that, the moderator index estimation/best combined moderator index construction would even extend to a mixed set of fully observed and partly observed moderators $\text{corr}(\Delta\text{Y}, \sum \tau_i W_i)$. For fully observed baseline moderator candidates we would use AW_i for W_i, whilst for partly observed moderator candidates we would use $W_{i1}$for W_i. The combination weights would come from a regression of ΔY on respective average or singleton moderator values.

In the case of selecting multiple IVs that are treatment moderators for mediation analysis, criterion (2) in Section 4.3.2.1 can be applied in combination with the moderation effect size developed in this section. More specifically, the best combined moderation effect size of multiple candidate IVs on the mediator should be no less than the best combined moderation effect size of these IVs on the outcome. In SPOKES, the interaction between treatment and baseline parental depression is an example of singleton fully observed candidate IV and the therapy groups in the treated arm dummy variables are examples of multiple partly observed candidate IVs. For the singleton R*parental depression candidate IV, $r(\Delta Y, AW)$ on both mediator and outcome are calculated and compared. For the therapy groups in the treated arm candidate IVs, $\text{corr}\left(\Delta\text{Y}, \sum \tau_q W_{q1}\right)$ on both mediator and outcome are calculated and compared. If a set of candidate IVs include both fully observed moderators (R*parental depression and R*parental education) and partly observed moderators (therapy groups in the treated arm), the combined moderator effect size, $\text{corr}(\Delta\text{Y}, \sum \tau_i W_i)$, are calculated using AW_p in place of W_i for the fully observed moderators, and using W_{q1} in place of W_i for the partly observed moderators.

4.3.3 Implementation of the combined IV-MI-BT approach

As the implantation steps of the MI-BT mediation approach have been described in Section 3.3.2 of Chapter 3, this section will focus on the implementation of the IV approach for mediation analysis and how does it combine with the MI-BT approach.

4.3.3.1 Programming steps for implementing the IV-MI-BT combined approach

Compared with the MI-BT approach, the IV-MI-BT combined approach includes an additional phase for deciding which variables to use as instruments for IV mediation analysis. In the IV-MI-BT procedure, the Baron and Kenny type mediation model for trials (plus measures confounders) will be replaced by an IV mediation model allowing for both measured and unmeasured confounding of the mediator-outcome relationship. Implementing the IV-MI-BT procedure requires a number of steps:

Phase 1: Specification of a list of potential IVs for mediation investigation and selection of the most promising ones to use in practice.

- **Step 1 – Specification of a list variables that could serve as IVs on theoretical grounds:** Based on the existing literature on treatment effect moderation in trials, and theoretical arguments about the suitability of interactions with randomisation as IVs for the mediator, a number of interaction term are selected (Section 4.3.2.1) as potential IVs.
- **Step 2 – Assessment of the impact of weak IVs:** Partial R-squared and F-statistics are calculated for each possible IV for each endogenous mediator in the single-mediator model. The methodological details of Shea's partial R-squared and F-statistics have been discussed in Section 4.2.1.3. These indices of the first stage model can be obtained by using the STATA command *ivreg2* under the *"first"* option. The F-statistics can be compared with the critical values tabulated in the Stock and Yogo (Stock and Yogo, 2002) paper to assess the relative bias to the bias of OLS estimator. Generally, the rule of thumb is that F-statistics needs to exceed 10 for a single endogenous variable and the relative bias from the 2SLS method is asymptotically approximately equal to $1/F$ (Sawa, 1969).
- **Step 3 – Comparison between moderation effect on parenting mediator and that on child outcome:** The moderation effects comparison approach proposed in Section

4.3.2.2 will be used to check that the IV set's effect on the mediator is no less than its effect on the child outcome.

- **Step 4 – IV orthogonalisation:** Orthogonalisation of IVs is achieved by regressing each individual IV on treatment group. The residual from this regression is then orthogonal to the randomised treatment group. In the IV analysis, the orthogonalised IVs (the residuals) will be used in the mediation model.

Phase 2: IV-MI-BT approach.

Similar to the MI-BT approach, the IV-MI-BT approach consists of two parts: the point estimate and the non-parametric BT inference. The point estimate of the casual parameters of interest is simply the MI-2SML estimator calculated from the original data using MICE to handle the missing data. The MICE procedure involving interaction terms as IVs will be introduced in the following section. The IV mediation model is listed in Equation 4-20 and Equation 4-21 and it also accounts for the hierarchical structure implied by the trial design. As introduced in Section 3.3.2.1 of Chapter 3, the non-parametric BT procedure for generating inferences consists of four steps. The IV-MI-BT follows the same procedure as the MI-BT approach. The only difference is that step 3 of the IV-MI-BT approach generates the MI-2SML estimator from the IV model. Thus the four steps of the IV-MI-BT approach are: Step 1 – BT resampling; Step 2 - Multiple Imputation; Step 3 - construction of MI-2SML estimator; Step 4 - bootstrap inferences. Again, the same measured confounder-selection procedure and the same effect standardisation approach as described in Chapter 3 will be applied to the IV-MI-BT mediation analysis. The bootstrap resampling strategy is determined by the trial design: thus, the bootstrap methods employed in IV-MI-BT or MI-BT to generate confidence intervals and p-values are identical for the same trial and I will re-use the methods described in Chapter 3. However, the MICE approach will be slightly different due to the involvement of IVs. The following section will provide details of the Multiple Imputation procedure for missing values in IV mediation analysis.

4.3.3.2 Relevant matters of MI with IVs for mediation analysis

Since the methods of Multiple Imputation for interaction terms and the general rules of constructing MI model for IV analysis have been reviewed in Section 4.2.3, I will focus on the implementation details of the MI for IV mediation analysis in this section.

Firstly, the variables included in the Multiple Imputations are all the variables included in the IV mediation analysis models and the auxiliary variables considered to be predictive of missingness. In the case of parenting programme RCTs in this project, the variables included in the IV models are: orthogonalised IVs, primary child outcome, all putative parenting practice mediators, intervention groups, selected baseline confounding variables (including the child outcome and the parenting mediator measured at baseline); auxiliary variables thought to be predictive of missingness in outcomes: non-primary child outcome measurement, measured parenting practices that were not considered as putative mediators, and measured baseline covariates that are not included in the analysis model.

Secondly, the imputation models needed to be modified compared to those described in Chapter 3. The exclusion restriction of IVs stipulates that there is no predictive effect of the IV on the outcome variable other than that operating via changing the mediator. Thus the IV is only included in the imputation model of its endogenous mediator, and not included in the imputation model of the child outcome. In this project, I investigate the putative mediators separately in a set of single mediator models and each endogenous mediator may have its only set of IVs. Thus in the imputation step, the IV will only be used to predict the missingness for its endogenous mediator but not for the other mediators.

In fact, the IVs included in the imputation model are the orthogonalised treatment by covariate interaction terms. As reviewed in Section 4.2.3, the JAV approach will be applied for IV imputation. Considering that the orthogonalised interaction terms (IVs) may not follow a normal distribution, the PMM technique (Chapter 3 Section 3.2.2) will be used to handle this. The imputation model and technique for the other variables with missing values are the same as the MI-BT approach. Again, I choose the number of imputations to be 20 based on the empirical results in Section 3.4.2 of Chapter 3.

4.4 Application of the IV-MI-BT combined approach to SPOKES

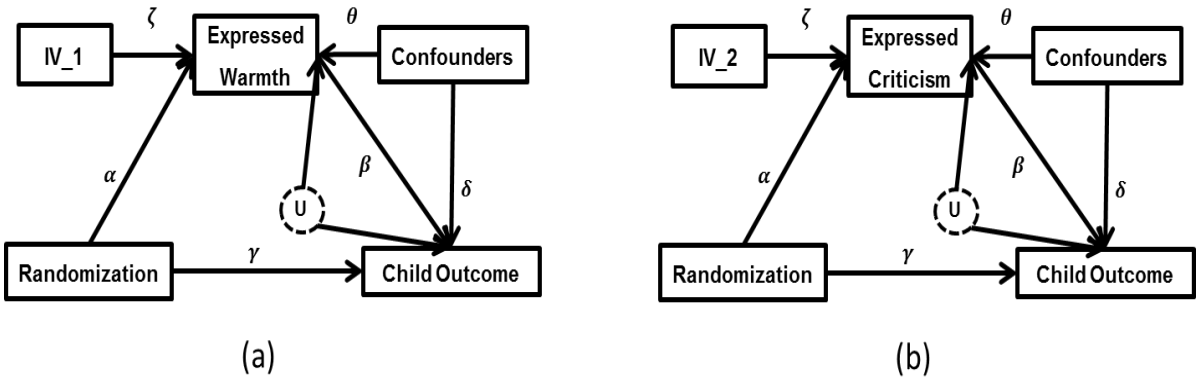
In this section, the IV-MI-BT approach will be applied to carry out mediation analyses for the same trial (SPOKES) for the purpose of comparing the results with those provided by the MI-BT approach. As shown in the previous chapter, the MI-BT approach estimated the causal mediation effects under the assumption of no unmeasured confounding of the mediator-outcome relationship and in the presence of missing data. To relax the assumption of no

unmeasured confounding, I will now estimate the causal parameters of interest for the same two mediators using the IV-MI-BT approach.

4.4.1 Single mediator IV mediation models

The analysis results of the MI-BT approach discussed in Chapter 3 suggested that two parenting practices mediators – expressed warmth and expressed criticism – mediated the treatment effect on child outcome. In this section, I will focus on assessing the causal mediation effects of these two putative mediators using an IV approach. The single mediator IV mediation models are illustrated in Figure 4-7.

Figure 4-7 Single-Mediator IV Mediation Models for SPOKES trial



The same set of measured confounding variables as in the MI-BT mediation model will be included in the IV mediation models. These variables are (1) child's gender, (2) child's reading ability, (3) parent's education, (4) parent's depression and (5) lone parent in addition to (6) child outcome measured at baseline and (7) parenting mediator measured at baseline.

4.4.2 Selecting instrumental variables for each putative mediator

Following the procedure specified in Section 4.3.3.1, firstly I carry on the phase I selection of the most promising IVs for mediation analysis for two putative mediators, expressed criticism and expressed warmth, respectively. Table 4-1 lists the F-statistics, partial and model R-squared indices, and sizes of moderation effects on the expressed criticism mediator and on the outcome for a list of potential instrumental variables.

The effect of potential multiple IVs on parenting and on child outcome are constructed using the Kraemer-type approach proposed in Section 4.3.2.2. The statistical indices are calculated

from two-stage IV linear regression mediation models with the measured confounders that are baseline child outcome, parenting mediator measured at baseline, parental education, parental depression, lone parent, child reading at baseline ability and child gender.

Table 4-1 Verification of instrumental variables for expressed criticism mediator

Potential IVs for expressed criticism mediator	**Shea's partial R-squared**	**1st stage model R-squared**	**F-statistics**	**IV effect on mediator**	**IV effect on outcome**
Treatment*lone parent	0.003	0.313	0.292	0.080	-0.092
Attendance (%) in the treated arm	0.003	0.313	0.283	0	0
Treatment*baseline expressed criticism	0.009	0.317	0.915	-0.111	-0.141
Treatment* baseline child outcome	0.012	0.32	1.251	0.240	-0.193
Treatment*parental depression	0.023	0.327	2.433	***-0.128***	***-0.033***
Treatment*parental education	0.028	0.331	2.953	***0.371***	***0.076***
Therapy groups in treated arm	0.114	0.39	0.970	***0.335***	***0.328***
Therapy groups in treated arm + Treatment*parental education + Treatment*parental depression	0.163	0.424	1.233	0.345	0.281

A set of IVs (shaded in grey) including therapy groups in the treated arm (9 binary variables), treatment effect by parental education interaction (1 binary variable), and treatment effect by parental depression interaction (1 binary variable) are selected for estimating the effect of expressed criticism mediator on child outcome. This set of IVs hold the biggest partial R-squares. The weak instrument critical values tabulated in (Stock and Yogo, 2002) paper requires that the F-statistic needs to be at least 4.8 for one endogenous variable with 11 IVs in order to reduce the bias to less than 30% of the OLS bias. Unfortunately, the F-statistics of 1.233 of selected IVs set indicates that this set of IVs is weak and the bias of the IV estimator is over 30% of the bias of the OLS estimator. On the other hand, the 30% plus bias relative to the OLS bias may not be unacceptable if the OLS bias is small. Additionally, each IV in the selected set (in **bold** and *italic*) and the combined set of IVs all have larger effects sizes for the mediator than the clinical outcome. As expected, the signs of the IV effects estimate on the mediator and on the outcome are consistent because the expressed criticism mediator and the outcome have a positive relationship. Thus both the magnitude and the direction of the IV effects would be consistent with the exclusion restrictions if direct and indirect IV effects were in the same direction.

Similarly, Table 4-2 lists the F-statistics, partial R-squared and first stage model R-squared, and moderation effects on the expressed warmth mediator and on the outcome for a list of potential instrumental variables. Following the same criteria for IV quality evaluation as for the expressed criticism, I selected the attendance (in percentage) of the training sessions of the parenting programmes in the treated arm as IV for the expressed warmth. The current F-statistic of the attendance variable is 3.117, which indicates roughly 70% bias reduction relative to the bias of the OLS estimator. The small F-statistics value also indicates that the attendance in the treated arm is a weak IV. The partial R-squared is very small (0.03), so that the variance inflation suffered by the respective IV estimator will be very large. Since the expressed warmth and the child outcome have a negative relationship, the signs of the IV on the mediator and outcome should be different. The results for the attendance IV showed that its effect on the mediator is larger than its effect on the outcome in absolute value with opposite direction as expected. However, the therapy group's effect on the mediator held the same sign as its effect on the outcome, which might indicate the existence of the path(s) between the IV and the outcome other than via the expressed warmth mediator. Thus, therapy groups in the treated arm are not selected.

Table 4-2 Verification of instrumental variables for expressed warmth mediator

Potential IVs for expressed warmth mediator	Shea's Partial R-squared	1st stage model R-squared	F-statistics	IV effect on mediator	IV effect on outcome
Treatment* baseline child outcome	0.001	0.176	0.088	0.04	-0.194
Treatment*baseline expressed warmth	0.002	0.177	0.171	-0.092	0.077
Treatment*parental education	0.003	0.178	0.317	0.085	0.141
Treatment*parental depression	0.003	0.177	0.262	0.004	-0.033
Treatment*lone parent	0.009	0.183	0.92	-0.249	-0.09
Attendance (%) in the treated arm	0.03	0.2	3.117	0.004	-0.001
Therapy groups in treated arm	0.158	0.306	1.415	0.426	0.333

The selected variables are orthogonalised to randomisation allocation by regressing each IV on the treatment group variable and the residual values form the orthogonalised IV. For the sake of simplicity, all the IVs mentioned in the following sections are the orthogonalised IVs.

4.4.3 SPOKES Mediation analysis using IV-MI-BT approach

So far, the two IV mediation models, one for expressed warmth and one for expressed criticism, have been set up properly including the same set of measured confounding variables, and different IV(s) for each endogenous mediator. The following section will demonstrate the results of causal mediation analyses using the IV-MI-BT approach. The bootstrap strategy is exactly the same as the MI-BT approach presented in Chapter 3. The number of Multiple Imputations is still 20 and the imputation models for the variables with missing values are similar to the MI model of the MI-BT approach. I list below the MI models updated due to the involvement of the IV.

a. The variables included in the imputation model of the expressed warmth parenting mediator:
    - measured baseline confounding variables (child's gender, child's reading ability, parental education, parental depression and lone parent)
    - measured baseline auxiliary variables (child's age, parental ethnicity, eligibility for free school meals)
    - treatment group randomisation assignment, school-year strata
    - expressed emotion variables measured at baseline and time point 2 (baseline expressed warmth, baseline expressed criticism, time point 2 expressed criticism)
    - child outcome measured at baseline and time point 2
    - positive parenting practices at time point 2 measured via different measurement methods (interview play, observed positivity, questionnaire positivity)
    - orthogonalised instrumental variable: attendance (%) of training sessions of parenting programme in the treated arm
b. The variables included in the imputation model of the expressed criticism parenting mediator:
    - measured baseline confounding variables (child's gender, child's reading ability, parental education, parental depression and lone parent)
    - measured baseline auxiliary variables (child's age, parental ethnicity, eligibility for free school meals)
    - treatment group randomisation assignment, school-year strata
    - expressed emotion variables measured at baseline and time point 2 (baseline expressed criticism, baseline expressed warmth, time point 2 expressed warmth)
    - child outcome measured at baseline and time point 2

- negative parenting practices at time point 2, measured by different measurement methods (interview smacking, observed negativity, questionnaire negativity)
- orthogonalised instrumental variables (therapy groups in the treated arm, treatment*parental depression interaction, treatment*parental education interaction)

c. The variables included in the imputation model for the orthogonalised treatment*parental depression interaction term:
   - measured baseline confounding variables (child's gender, child's reading ability, parental education and lone parent)
   - measured baseline auxiliary variables (child's age, parental ethnicity, eligibility for free school meals)
   - school-year strata, child outcome measured at baseline, expressed criticism measured at baseline and time point 2

d. The variables included in the imputation model for the orthogonalised treatment*parental education interaction term:
   - measured baseline confounding variables (child's gender, child's reading ability, parental depression and lone parent)
   - measured baseline auxiliary variables (child's age, parental ethnicity, eligibility for free school meals)
   - school-year strata, child outcome measured at baseline, expressed criticism measured at baseline and time point 2

Since the therapy groups in the treated arm are considered as IVs for the expressed criticism, they are not included in the imputation models of the other variables with missing values. No missing values were found in the variables "attendance (%) of training sessions" and "therapy groups in the treated arm".

The mediation analysis results for expressed warmth and the expressed criticism mediator using the IV-MI-BT approach are listed in Table 4-3. The estimates of the causal mediation effects are standardised according to the standardisation method introduced in Section 3.3.2.4 of Chapter 3.

Table 4-3 Results of mediation analysis for expressed warmth and expressed criticism mediators using the IV-MI-BT method

Putative Mediator	Causal mediation parameter	Estimate	SE	P-value	Bias Corrected 95% BT CI
Warmth	α	0.46	0.33	0.17	(-0.12, 1.22)
	β	-0.33	2.92	0.28	(-0.83, 2.6)
	γ	-0.38	2.61	0.11	(-0.92, 0.18)
	$\alpha\beta$	-0.15	2.6	0.35	(-2.62, 0.12)
	$\gamma + \alpha\beta$	-0.54	0.12	<0.01	(-0.73, -0.28)*
Criticism	α	-0.39	0.21	0.11	(-0.68, 0.16)
	β	0.32	0.16	0.08	(-0.07, 0.61)
	γ	-0.4	0.14	0.02	(-0.63, -0.12)*
	$\alpha\beta$	-0.12	0.13	0.27	(-0.3, 0.09)
	$\gamma + \alpha\beta$	-0.52	0.14	<0.01	(-0.79, -0.2)*

In this table, α is the effect of the treatment on the mediator (ETM), β is the effect of the mediator on the outcome (EMO), $\alpha\beta$ is the indirect treatment effect (IE), γ is the direct treatment effect (DE) and $\gamma + \alpha\beta$ is the total treatment effect (TE). The results of the expressed warmth IV mediation analysis show that the SPOKES parenting intervention reduced child antisocial behaviour problems by 0.54 standard deviations, of which 0.15 standard deviations are due to increasing parental expressed warmth (27.8% of the total effect) and 0.38 standard deviation are due to factors other than expressed warmth. The results of the expressed criticism IV mediation analysis show that the SPOKES parenting intervention reduced child antisocial behaviour by 0.52 standard deviations, of which 0.12 standard deviations are due to reducing parental expressed criticism (23.1% of the total effect) and 0.4 standardised deviations are due to factors other than expressed criticism. Although the magnitude of the causal direct and indirect effects are large/moderate, they are not statistically significant at 5% level based on the bias correct confidence interval (0 is included in the interval of the direct and indirect effects for both putative mediators) due to variance inflation caused by weak IVs.

4.4.4 Comparing the analysis results of IV-MI-BT and MI-BT approaches

To further study the impact of the two mediation analysis methods, I compare the mediation analysis results of the IV-MI-BT approach with that of the MI-BT approach in this section. Slightly different to the MI-BT approach presented in Chapter 3, the MI-BT mediation approach used here includes the IVs (treatment*covariates interaction terms) as mediator-

outcome confounders in the analysis model for the purpose of direct comparison of methods. Table 4-4 lists the mediation analysis results of the same two mediators (expressed warmth and criticism) using MI-BT approach.

Table 4-4 Results of mediation analysis for expressed warmth and expressed criticism mediators using the MI-BT method including interaction terms as covariates

Putative Mediator	Causal mediation parameter	Estimate	SE	P-value	Bias Corrected 95% BT CI
Warmth	α	0.4	0.31	0.2	(-0.07, 1.16)
	β	-0.23	0.07	0.01	(-0.36, -0.09)
	γ	-0.38	0.16	0.02	(-0.6, 0.06)
	$\alpha\beta$	-0.09	0.07	0.16	(-0.27, 0.02)
	$\gamma + \alpha\beta$	-0.47	0.14	0.01	(-0.71, -0.16)
Criticism	α	-0.38	0.18	0.03	(-0.74, -0.02)
	β	0.42	0.12	0.01	(0.21, 0.66)
	γ	-0.32	0.17	0.15	(-0.65, -0.13)
	$\alpha\beta$	-0.16	0.08	0.05	(-0.39, -0.03)
	$\gamma + \alpha\beta$	-0.47	0.16	0.03	(-0.8, -0.34)

As expected, comparison of Table 4-3 and Table 4-4 shows that the IV-MI-BT and the MI-BT approaches provide similar estimates and standard errors of the TE ($\gamma + \alpha\beta$) and of the ETM (α) in both warmth and criticism mediation models. We know that the differences between the two approaches mainly lie in the different estimation of the effect of EMO (β) under different assumptions. If the estimate of β is different, then the estimates of the IE ($\alpha\beta$) and the DE (γ) are expected to be different. Thus I will focus on the comparison of the estimates of β between the two approaches. The β estimate of expressed warmth is -0.23 with a standard error equal to 0.07 in the MI-BT approach, whilst the IV-MI-BT approach gives an estimate of -0.33 with a standard error equal to 2.92. Under the assumption of the IV method, the difference of the estimates of the two approaches is attributed to the unmeasured confounding of the warmth-child outcome relationship. However, the instrument of warmth holds weak explanatory power (partial R-squared = 0.03 in Table 4-2), so that it causes large standard error inflation. For expressed criticism, the β estimate and standard error provided by the MI-BT and IV-MI-BT approaches are 0.42 (0.12) and 0.32 (0.16) respectively. This indicates that after allowing for unmeasured confounding between criticism and child outcome, the procedure obtains a slightly smaller estimate of the effect of criticism on child outcome. The standard error increased by 33%, which is relatively small in

magnitude compared with the variance inflation in expressed warmth. This may be due to the stronger explanatory power of the IVs (partial R-squared = 0.163 in Table 4-1). Both IVs and the included exogenous variables are predictors in the first stage model and the model R-squared is 0.42 (see Table 4-1), which indicates a moderate-large explanatory power. Thus the variance inflation of the MI-2SML estimator of β in the expressed criticism mediation model is not excessive.

4.5 Discussion

4.5.1 The strengths of the IV-MI-BT approach for mediation analysis

Similar to the MI-BT approach, the IV-MI-BT approach is a practical and flexible method for addressing both measured and unmeasured confounding of mediation analysis in the presence of missing data. As the strengths of the MI-BT approach have been discussed in Chapter 3, I will focus on the properties of the IV-MI-BT approach related to the IVs. Most of the IV analyses ignore missing values when use standard software, whilst the combination of IV and MICE sorts out the missing data issue. The flexibility of MICE allows various types of IVs, including interaction terms. Although the current IV analysis is based on a two-stage ML-type estimator, the IV-MI-BT approach can be used in conjunction with causal effects estimators that have favourable properties, such as the limited-information maximum likelihood estimator for weak instruments.

4.5.2 Limitations of the IV-MI-BT approach

As described in this chapter, The IV 2SML estimator proposed in the IV-MI-BT approach is a consistent estimator under certain assumptions. Similar to the ML estimators, the IV 2SML estimator's asymptotic property requires a large sample size. In addition to the asymptotic property, this IV-based estimator will be imprecise (large standard error) and biased when sample size is small. Thus, large sample size is a critical condition for the precision and accuracy of the IV 2SML estimator. However, the sample size of single trials is often small. Thus combining data from multiple trials of similar studies may be a good idea to improve the estimates via increasing the sample size. Chapter 5 of this thesis will undertake further explorations inspired by this idea. Additionally, Bound et al. (Bound et al., 1995) have shown that if the IV is weakly related to the endogenous variable then the bias of the IV estimator can be substantial under slight violation of the IV exogeneity assumption: therefore, it is advisable to use an IV that is strongly correlated with the endogenous variable. However, in

practice, this is often difficult to achieve. Moreover, Martens and others (Martens et al., 2006) have shown that if strong hidden confounding is to be expected and an IV has been used that is moderately or strongly related to the endogenous variable, it is likely that the IV assumptions are violated (the IV might be related to the strong hidden confounding), resulting in a biased effect estimate. Therefore, in practice, IV methods are more appropriate in the case of moderate confounding, as strong instruments cannot be found and assumptions will be easily violated.

4.5.3 Estimators for weak instruments

It has been shown that the finite sample distribution of the IV estimator can depart dramatically from the asymptotic normal distribution under weak instruments and the bias is in the direction of the OLS estimator. To handle weak instruments in the linear IV model, more robust estimators with weak instruments were introduced for the linear IV model with one endogenous explanatory variable by Stock *et al* (Stock et al., 2002). These robust estimators include the *limited-information maximum likelihood* (LIML) estimator (Rothenberg, 2007), *bias adjusted two stage least square* (BTSLS) estimator (Donald and Newey, 2001), the *Jackknife instrumental variables estimator* (JIVE) (Angrist et al., 1995) and the *Fuller-*κ estimator (Fuller, 1977, Hahn et al., 2004). The LIML estimator is computed with a little more effort than the 2SLS estimator, and the asymptotic properties of the 2SLS and LIML estimators are the same if the instruments are not too weak (Davidson and Machkinnon, 1993). However, these estimators were all developed without considering missing values and multi-level data structures. Moreover, most of the analyses in the weak instruments literature are conditional on IV exclusion restriction. Failure of the exclusion restriction, particularly in combination with weak instruments, leads to additional complications (Hahn and Hausman, 2003). The challenges of weak instruments and the violation of exclusion restriction require further development of the IV-MI-BT approach.

Chapter 5 Causal Individual Participant Data (IPD) Meta-Mediation Analysis

5.1 Introduction

It is well known that parenting programmes are the most effective intervention to change persistent child antisocial behaviour (NICE, 2013). Multiple RCTs of parenting intervention have been conducted for preventing and reducing child antisocial behaviour. As reviewed in Chapter 1, even though the effect of parenting intervention has been very well established, mediation analysis to understand the mechanism of the intervention is rarely done. The three trials of IY parenting programmes (SPOKES, CPT and HCA) introduced in Chapter 2 of this thesis provide a rich data source to investigate novel research questions related to mechanisms and to provide synthesised estimates of the effects of interest. Thus, in this chapter, I propose to pool data from multiple trials of parenting intervention for the purpose of conducting a meta-mediation analysis. *Meta-analysis* refers to statistical methods for combining and analysing quantitative evidence from multiple related studies to produce results based on a whole body of research (Riley et al., 2010). I refer to a meta-analysis for the purpose of investigating mediation of a treatment effect as *meta-mediation analysis*.

5.1.1 Reasons for pooling data in causal meta-mediation analysis

The objective of mediation analysis is to answer the question regarding how treatment changes the outcome. Pooling data from multiple trials of the same intervention provides the opportunity to investigate further interesting research questions such as: How do mediation effects of interest vary according to trial? Does the between-trial heterogeneity of various effects support the hypothesised mediators as mechanisms? Can the mediation effects of multiple trials be synthesised to provide more efficient estimates? In the following paragraphs, I will explain these three questions in detail.

Analyses of the between-trial variability in mediation effects can provide a better understanding of the total intervention effects in different trials. In general, if there is between-trial variability in the total effect of the intervention, then we need to provide a separate estimate of the intervention effect for each trial. It is interesting to understand the sources of such intervention effect heterogeneity. For example, we might ask whether a reduced total effect in one trial can be explained by the specific intervention implemented

and delivered in that trial having less of a benefit in terms of the target intermediate variable. In the case of parenting intervention mediation analysis using pooled data, I am thus interested in knowing how the effect of the intervention on parenting practice (α), the effect of parenting practice on child outcome (β), and the direct effect of the intervention on child outcome (γ) varies with each trial. In other words, I am interested in investigating whether there is an interaction between these effects and the trial.

In addition, findings from such interaction analyses may provide further empirical support for putative mediators as explanations of the mechanisms. For example, provided that we are looking at samples from the same target population, we expect the same mechanisms to operate and determine the outcome. Thus, in the context of parenting trials with putative parenting practice mediator variables, we are expecting their effects on child outcome (β) to be constant across trials. A non-significant between-trial heterogeneity test for this parameter would be consistent with a mechanism hypothesis.

A final benefit of pooled data meta-mediation analysis is that it can provide more efficient estimates of mediation effects if some of the analysis model parameters can be held constant across trials (no between-trial heterogeneity). Simply put, estimating single β using data from multiple studies is more efficient than estimating distinct β_i using data from a corresponding i study. This is because under effect homogeneity, a larger and more representative sample can contribute to the estimation of the parameters of interest. This potential precision improvement is crucial for the IV-MI-BT mediation analysis proposed in the previous chapter. Briefly, as discussed in Chapter 4, an instrumental variable approach requires a large sample size to reduce variance inflation and finite sample bias. The application in Chapter 4 showed that the IV-MI-BT estimate of the causal indirect effect via expressed criticism was insignificant at the 5% level with a wider 95% confidence interval, whilst the MI-BT estimate was significant at the 5% level with a narrower confidence interval. This is a good example of the variance inflation suffered by IV estimators and leading to loss of power. It has been shown that tests of indirect intervention effects generally require larger sample sizes than the primary test of the total treatment effect. Simulation results suggest that the sample size needs to be around four hundred to detect an indirect effect including small α or β (size$= 0.14$) with 0.8 power using a traditional linear

regression approach (Fritz and MacKinnon, 2007). Moreover, in cases of clustered data, a larger sample size will be required to account for the intra-class correlation.

However, the sample sizes of parenting programmes' RCTs are rarely over two hundred, and usually less, as the trials were initially powered for testing the total treatment effect on child outcome. Thus, there is a need to generate large data for mediation analysis of parenting programmes. Although the sample size of a single RCT of a parenting programme is small, multiple trials of the same intervention with similar trial designs and measures could usefully be combined. Therefore, combining the individual-participant data of these trials is one way to achieve a large sample size and potential power improvement in the mediation analysis.

In this chapter, the data from three trials of IY parenting programmes (SPOKES, CPT and HCA) will be pooled together for a meta-mediation analysis using the IV-MI-BT approach to investigate the synthesised mediation effects and potentially regain precision compared with the single trial IV mediation analysis in Chapter 4.

5.1.2 Statistical challenges of mediation analyses using pooled data

Mediation analysis of a pooled data set is not a straightforward application of the mediation analysis of the single trial using the combined data. Compared with the analysis of a single trial, meta-mediation analysis faces several new challenges:

Firstly, the trials might have been conducted with different populations, so that we need to find a way to take the effects of subpopulations into account in the meta-mediation analysis. As mentioned in Chapter 2, the participants in the three trials of this project (SPOKES, CPT and HCA) were selected using slightly different inclusion and exclusion criteria. For each trial, the participants are considered as a representative sample of their target population. The target populations of the three parenting trials are different in terms of child's age, antisocial behaviour severity, and geographical locations (see Chapter 2, Table 2-1). Therefore, it is possible that different trials have different expected values for the child outcomes and the parenting mediators due to population differences.

Secondly, even though the trials included in the meta-mediation analysis all used parallel group designs, design variations are quite common. For instance, trial designs might have

been changed based on the experiences of previous studies or adapted to real-world situations. These differences in trial design should be modelled appropriately in the pooled data analysis. In this project, SPOKES is an RCT of a combined IY plus literacy parenting programme versus telephone helpline control and employed stratified randomisation, CPT is a cluster quasi-randomisation study of IY only parenting programme versus waiting list control, and HCA employed a factorial trial design (four trial arms: IY only, literacy only, IY and literacy combined, and the service as usual/'signposting' control group) and participants were randomised to different intervention groups in different recruitment cohorts (strata). Thus ideally, different ingredients of the interventions and control conditions should be considered in the analysis model. Additionally, the hierarchical data structure implied by each trial design should also be modelled accordingly.

Thirdly, even if the trials implemented the same active and control conditions and sampled participants from the same target population, we would expect to observe variability in the estimates of causal parameters of interest (including the confounding effects). We therefore seek to empirically distinguish between effect heterogeneity arising "by chance" and due to population parameters varying over trials. The results from such tests can then help us to explain how intervention effects are generated across a range of trials (with possibly different treatment implementations/target populations). Statistically, we are seeking to test interactions between the effects of interest and the trials. In the single trial mediation analyses in the previous two chapters, I estimated the direct effect of the treatment on the outcome (γ), the effect of the treatment on the mediator (α), the effect of the mediator on the outcome (β), and the effects of the confounders on the mediator ($\boldsymbol{\delta}$) and on the outcome ($\boldsymbol{\theta}$). In the pooled data mediation analysis, the existence of an interaction between each of these effects and the trials needs to be tested. Theory would suggest that parameters representing mechanistic effects in the same target population, such as the effect of the mediator on outcome (β) or effects of confounders ($\boldsymbol{\delta}$ or $\boldsymbol{\theta}$), do not differ between trials. However, there is still a need to seek empirical support for this assumption of no between-trial heterogeneity of β, especially when trial target populations vary.

Testing effect-trial interactions with multiple trials provides us a technical challenge. The mediation analysis approaches proposed in the previous two chapters calculate confidence intervals of the effect estimate using a non-parametric bootstrap approach that mimics the

trial design. The meta-mediation analysis involves multiple trials (e.g. trials A, B, and C) with different trial designs and requires testing of composite hypotheses (e.g. $\beta^A = \beta^B = \beta^C$). To meet these requirements, the bootstrapping strategy will be further developed to respect trial-varying designs and to generate tests of composite hypotheses.

5.1.3 Chapter outline

This chapter is organised in two main parts: Development of meta-mediation analysis methodology (Part I) and application of the methods to meta-mediation analysis of a pooled data set consisting of data from the SPOKES, CPT and HCA trials (Part II). Specifically, I will proceed as follows: The next Section 5.2 reviews existing statistical methodologies that are helpful for conceptualising and addressing the challenges posed by meta-mediation analysis. Section 5.3 introduces a systematic approach to meta-mediation modelling of pooled data, identifying necessary trial interactions and selecting a final analysis model. This includes a novel non-parametric bootstrap approach for testing between-trial heterogeneity of the effects of interest. The bootstrap strategy and the MI approach of the IV-MI-BT method will be adapted for meta-mediation analysis and will form the IV-MI-BT meta-mediation method. In Part II of this chapter, this new IV-MI-BT meta-mediation analysis method will be applied to analyse pooled data from three parenting trials (SPOKES, CPT and HCA). The meta-analysis results of parenting programmes on child antisocial behaviour will be reported and interpreted in Section 5.4. Briefly, I found statistically significant total effects and direct effects of the parenting interventions on the child outcome. However, the synthesised estimates of the indirect effect via parental expressed warmth or parental expressed criticism are small and failed to reach significance at the 5% level. The precision of the synthesised estimate of causal effects of interest is improved compared with the single trial estimate. Finally, Section 5.5 discusses the strengths and limitations of the IV-MI-BT meta-mediation analysis.

5.2 Review of related statistical methodology

5.2.1 Individual participant data meta-analysis and aggregate data meta-analysis

The use of meta-analysis for assessing the potential benefits of health care interventions has greatly increased over the years (Lambert et al., 2002, Sutton and Higgins, 2008, Riley et al., 2010). The main motivation of meta-analysis is to combine information in order to increase the precision of statistical inferences. Conventional methods for meta-analysis synthesise

aggregate *study-level* data obtained from study publications, such as treatment effect estimates and their associated uncertainties. One might consider this approach an *aggregate data (AD) meta-analysis*. An alternative but increasingly popular approach is *individual participant data (IPD) meta-analysis* or *meta-analysis of individual patient data* or *integrative data analysis (IDA)*, in which the raw individual-level data of each study are pooled together and included in the synthesis. The term *individual participant data (IPD)* relates to the data recorded for each participant in a study. On the contrary, the term *aggregate data (AD)* relates to information averaged or estimated across all individuals in a study. Such aggregate data are derived from the individual participant data themselves and the individual participant data can be considered as the original source material.

The AD meta-analysis approach bases its findings on the reported summary data that are often derived, presented and analysed differently across studies. For example, some studies may have used the transformed and standardised measurement, whilst others may have used the original measurement. In another case, some studies may have reported the association of two variables conditional on confounding variables of these two, whilst others may have reported the association without adjusting confounding factors. The AD meta-analysis relies on the reported summary data. In the above circumstances, we may not be able to get the same measure of interest across studies. In addition, the fact that statistically or clinically significant results are more likely to be reported leads to selective reporting within and across studies and amplifies publication bias. For example, the Depression Anxiety Stress Scale 21 (DASS-21: (Henry and Crawford, 2005) measures mental health in three dimensions: depression, anxiety and stress. However, some studies that used the same measurement instrument (DASS-21) might have reported only the statistically significant anxiety score. If our measure of interest is the depression score measured using DASS-21 in the meta-analysis, the AD meta-analysis excludes these studies due to the unavailability of the depression score, which can lead to publication bias in the synthesised estimate. Individual participant data include information that was not reported in the original study, avoiding publication bias to a certain extent. The drawbacks listed above limit the applicability of AD meta-analysis for synthesising the aggregated data in a consistent and meaningful way.

Another limitation of AD meta-analysis is that it almost restricts the topics of inquiry to those questions that have already been addressed within individual studies or to the investigation of study level characteristics. For example, the literature on parenting programme RCTs mainly focuses on testing the effectiveness of parenting interventions for improving child antisocial behaviour outcomes, but the mediation of the intervention effect via targeted parenting practices is not often conducted. In this case, the AD meta-analysis method is not applicable to provide a synthesised mediation effect estimate using parenting RCTs that only report the effect of intervention on the parenting practices and the effect of intervention on child outcomes and do not estimate the effect of the parenting mediator on the child outcome. Most importantly, AD meta-analysis has no access to the within-trial variability, which precludes us from optimally modelling the data and generating more accurate estimates. Thus, the pooling of original data from contributing trials together to conduct an IPD meta-analysis is proposed in this project.

The IPD meta-analysis is acknowledged as the gold standard methodology for carrying out a meta-analysis (Stewart and Parmar, 1993). This is because IPD meta-analysis has a list of advantages compared with AD meta-analysis. Since IPD meta-analysis includes the original data, it can be independent of the objective, testing, significance and reporting of the contributing studies that were published. The IPD meta-analysis can go beyond the "grand mean" (Smith et al., 1997). More specifically, missing data can be recognised and accounted for at the individual level; results for unpublished data (obtained or estimated) can be incorporated for the purpose of reducing publication bias; the desired information can be directly derived and standardised from the raw data; the analysis method can also be standardised across all studies, including adjusting the same confounding factors to obtain interpretable synthesised results. Importantly, the investigation of novel research questions that were not considered in the original studies, such as intervention mechanisms (mediation analysis), is also facilitated by analysing the individual participant data. Additionally, individual participant data enables complex statistical modelling of between- as well as within-trial variability, such as modelling the between-trial heterogeneity of the effects of interest and the within-trial clustered (and hierarchical) data structure.

5.2.2 Statistical methods for individual participant data (IPD) meta-analysis

The pooled individual participant data from multiple trials might be clustered data because outcomes from participants within trials are likely to be correlated. The analysis of IPD should preserve such trial clusters. Trial clusters can be retained during analysis by using a two-step or a one-step approach. A two-step approach is so named because it analyses the pooled data in two steps. In the first step, the IPD of each trial is analysed to generate relevant trial-level statistics by using a statistical model that is appropriate for the type of data. In the second step, these summary results are combined across trials using the AD meta-analysis method. The two-step approach can be considered as an improved AD approach. A one-step approach analyses all the IPD simultaneously while accounting for trial clustering. Both one-step and two-step IPD meta-analyses are facilitated by the flexible multilevel modelling approach (Turner et al., 2000). If pooled average effects are of interest, then the two-step and one-step approaches will produce identical results (Stewart et al., 2012). However, if the exploration of participant-level covariates is of interest (e.g. treatment by covariate interactions), then the one-step IPD meta-analysis approach has better power (Simmonds et al., 2005, Riley et al., 2010). In this project, the effects of interest are more complex than simple mean effects (including the products of effects). I also wish to adjust patient level confounding factors, which can increase the power of the analysis to detect mediation effects of interest. Thus the one-step IPD meta-analysis approach will be applied. Since only the one-step approach will be applied in this project, I will focus my review on multilevel modelling for one-step IPD meta-analysis. For the sake of simplicity, the term *IPD meta-analysis* used in subsequent paragraphs indicates *one-step IPD meta-analysis*.

The IPD meta-analysis of a pooled trials data involves a series of considerations. The first one is the choice between *fixed* trial effects and *random* trial effects. In a simple case of analysing pooled data from RCTs assessing the treatment effect on a continuous outcome, one can fit a linear regression model using pooled IPD with fixed trial effects to allow the outcomes to differ across trials. Such a fixed trial effect may be implemented by adding a set of dummy variables of the trial memberships in the regression model. The fixed effect approach can also allow the treatment effect to vary with trial via including interaction with trial dummy variables. However, this approach estimates a nuisance parameter for every trial included and separate treatment effects for different trials, so such an analysis is equivalent to separate analysis of each trial and no power gain is achieved. An alternative

approach is to regard the trial effects on the outcome as random effects. A random effects meta-analysis model may include random effects u_j of trial on outcome as well as the effects v_j of trial on treatment effect. This random effect model allows the deviation of each trial's true treatment effect from the average (trial effect as random intercept and treatment group as random slope).

$$Y_{ij} = \beta_0 + \beta_1 T_{ij} + u_j + v_j T_{ij} + \epsilon_{ij} \qquad \text{Equation 5-1}$$

where Y_{ij} is the outcome of the i^{th} individual in the j^{th} trial, T_{ij} is the treatment group indicator for each individual participant i of trial j, β_0 is the intercept, β_1 is the coefficient of T_{ij}, and ϵ_{ij} represents the residual error. A random effects formulation considers the trial-level random effects to represent samples from a larger population of possible random effects and thus uses fewer variance parameters to describe the data. Unless one is specifically interested in treatment effects under a specific setting, the random effects formulation is conceptually more appealing due to its generalizability, with β_1 being interpreted as the average treatment effects across all possible trials. However, a reasonable number of replicate trials are required to estimate the variances of the random effects and this excludes the use of random effects for IPD meta-analysis with only a few (say less than 10) contributing studies. As only three parenting programme trials are included in this project's IPD meta-analysis, the trial effects will need to be considered as fixed effects in the analysis model for this pooled data.

The between-trial heterogeneity in the effects of covariates (including those of treatment and confounding variables) is an important aspect of IPD meta-analysis that must be considered. When trial effects are included in the analysis model as fixed effects, evaluating the existence of between-trial heterogeneity in the effects of covariates is in fact testing whether the effects of the covariates vary with trial, or, say, testing the trial-covariates interactions. If the effects are constant across trials (no existence of trial-covariate interaction), then an estimate of the overall effect using pooled data provides a more precise inference compared with any effect estimate derived from a single trial. On the other hand, the existence of trial-covariate interaction will provide an insight into the effects of interest across trials.

Between-trial heterogeneity in the effects of covariates may be evaluated by testing hypotheses involving multiple parameters – so-called *composite hypotheses*. All F-tests, likelihood ratio tests and bootstrapping approaches have been used for this purpose. Here, I prefer to use the (non-parametric) bootstrapping method because it relaxes the distributional assumptions of the data and also it follows the same framework as the statistical inference generation process proposed in the previous chapters. In the following section, I will review the related methodology of the bootstrapping method for testing hypotheses involving multiple parameters.

5.2.3 Bootstrap tests of composite hypotheses

The bootstrap pivot approach for testing a hypothesis about a single parameter has been reviewed in Section 3.2.4.4 of Chapter 3. The bootstrapping approach for testing a hypothesis involving multiple parameters is a natural extension of the single parameter approach. Correspondingly, a *confidence region* is a multi-dimensional (q-dimensions) generalization of a *confidence interval*. It is a set of points in a q-dimensional space, often represented as an ellipsoid around a point which is an estimated solution to a problem. In the case of two parameters (two dimensions), a *confidence ellipsoid* is defined broadly as an ellipse-shaped joint $100(1-\alpha)\%$ confidence region (Alexandersson, 2004).

A confidence ellipsoid can be used to test a composite hypothesis H_0 for a vector parameter $\phi = \left(\phi_1,\ \phi_2,\ \ldots, \phi_q\right)^T$ (Scheffe, 1999).

$$H_0\colon\ \phi_1 = \ \phi_2 = \ \phi_3 = \cdots = \ \phi_q = 0 \qquad \text{Equation 5-2}$$

where ϕ_i are q linearly independent estimable functions. H_0 is rejected if and only if the $100(1-\alpha)\%$ confidence ellipsoid fails to cover the hypothesised parameter vector $(\phi_1,\ \phi_2,\ \ldots, \phi_q) = (0,\ 0,\ \ldots, 0)$. In the case of multiple linear regression of a response variable (Y) on a set of q explanatory variables $(\mathbf{X})$ with normally distributed errors, the commonly used method for testing the hypothesis that the slope coefficients $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$ are all equal to particular values $\boldsymbol{\beta}^{(\mathbf{0})}$ is the F-test. Alternatively, the hypothesis can be tested via the $100(1-\alpha)\%$ joint confidence region for (Monette, 1990):

$$\Pr\left[\frac{\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(\mathbf{0})}\right)'\mathbf{V}^{-1}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(\mathbf{0})}\right)}{qS_{\varepsilon}^{2}} \leq F_{\alpha,\ (q,\ n-q-1)}\right]=1-\alpha \qquad \text{Equation 5-3}$$

where $\mathbf{V}$ represents the square submatrix consisting of the entries in the q rows and q columns of $(\mathbf{X}'\mathbf{X})^{-1}$ for the slope coefficients in $\widehat{\boldsymbol{\beta}}$, S_{ε}^{2} is the mean square error, $F_{\alpha,\ (q,\ n-q-1)}$ is the critical value of F with q and $(n-q-1)$ degrees of freedom, corresponding to a right-tail probability of α. In this case, ϕ_i in Equation 5-2 is parameterised as $\widehat{\beta}_i-\beta_i^{(0)}$. The joint confidence region for $\boldsymbol{\beta}$ is thus all $\boldsymbol{\beta}$ for which $\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(\mathbf{0})}\right)'\mathbf{V}^{-1}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^{(\mathbf{0})}\right) \leq qS_{\varepsilon}^{2}F_{\alpha,\ (q,\ n-q-1)}$. This region represents an ellipsoid in the q dimensional parameter space of the slope coefficients. $\widehat{\beta}_i$ is assumed to be approximately normal, and the distribution $q*F(q,n-q-1)$ approaches a χ^2 distribution with $(q-1)$ degrees of freedom when the sample size n is large. However, as in the scalar case, such distributional approximations will often be unreliable (Davison and Hinkley, 1997) p.233). In addition, the normality assumption of response variable Y may not hold in some cases.

To relax the distributional assumption required by the test, a non-parametric bootstrap approach can be applied to approximate the distribution of the test statistics under the composite null hypothesis. As an analogue to the bootstrap pivot approach introduced in Section 3.2.4.4, the bootstrap p-value for hypothesis testing of a vector parameter H_0 (Equation 5-2) can be calculated as

$$p=\Pr(D^{*2}\geq d^{2}|\hat{F}) \qquad \text{Equation 5-4}$$

where D^* is the pivot statistic calculated for each of B bootstrap samples, d is calculated using the observed sample, and $\hat{F}$ is bootstrap sampling distribution of the parameter estimate $\widehat{\boldsymbol{\phi}}$. If we had the bootstrap estimated p quantiles a_p^*, then the $100(1-\alpha)\%$ joint confidence region for $\boldsymbol{\phi}$ would be the region which $D\leq a_{1-\alpha}^{*}$. The pivot statistic D is a function of $\boldsymbol{\phi}$, which is a multi-dimensional generalisation of $\boldsymbol{\phi}$ and it calibrates the multivariate data into univariate data. However, there is no vector analogue of the adjusted percentile methods (Davison and Hinkley, 1997).

5.2.4 Mahalanobis distance for multi-dimensional points to the centre of mass

Mahalanobis distance (MD) measures the generalised distance of a multi-dimensional point from the centre of mass (De Maesschalck et al., 2000). Thus it is a good candidate to serve as the pivot statistic of a vector of parameters involved in a composite hypothesis. For an observation $\mathbf{x} = \left(x_1, x_2, \dots, x_q\right)^T$ from a group of observations with mean $\boldsymbol{\mu} = \left(\mu_1, \mu_2, \dots, \mu_q,\right)^T$ and data variance-covariance matrix $\mathbf{V}$, the DM of the observation is calculated as

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^{\mathbf{T}} \mathbf{V}^{-\mathbf{1}} (\mathbf{x} - \boldsymbol{\mu})} \qquad \text{Equation 5-5}$$

This definition shows that in the original variable space, the MD takes into account the correlation in the data because it is calculated using the inverse of the variance–covariance matrix of the data set of interest. In fact, MD calibrates the multivariate data into univariate data and reduces the dimensionality without making distributional assumptions. Given the above properties of MD, it is good candidate as the pivot statistic for the bootstrap approach reviewed in Section 5.2.3. A novel composite hypotheses testing approach combining the non-parametric bootstrap approach with MD is developed in the following section.

5.3 A novel IPD meta-mediation analysis approach

Even though framework and statistical methods for analysing IPD have been of great interest to many researchers in recent decades (Sutton and Higgins, 2008), statistical methods for meta-analysis in the field of causal mediation are rarely investigated. To fill this gap, this section will introduce a systematic approach for selecting an appropriate analysis model for IPD meta-mediation analysis, ensuring that this analysis model reflects the designs of contributing trials. Following the same strategy for generating statistical inferences as in previous chapters, a non-parametric bootstrap approach for meta-mediation analysis will be developed, in combination with Mahalanobis distance. As missing data is a common issue for effects estimation conditioning on multiple confounding variables, MI will be applied within each trial contributing to the IPD meta-mediation analysis.

5.3.1 A systematic approach to IPD meta-mediation modelling

The systematic approach to meta-mediation modelling proposed in this section addresses the statistical challenges brought out in Section 5.1.2 for IPD meta-mediation analysis. This section consists of two parts. It begins by introducing principles for constructing a full meta-mediation analysis model which allows all model parameters to vary with trial. This is then followed by a novel bootstrap procedure for model simplification - testing whether restrictions can be imposed to hold some parameters constant across trials.

5.3.1.1 Principles for constructing the full IPD meta-mediation model

I apply three principles for parameterisation of the mediation model for pooled trials data:
Firstly, the trial membership variables need to be included in the IPD meta-mediation analysis model because the populations may be different across contributing trials (see Section 5.1.2). The effects of trial can be considered as random effects or fixed effects under different assumptions and for different data (see review of IPD meta-analysis methods in Section 5.2.2). In this project, I will focus on the situation when the number of the contributing trials is very small (less than 10) so that it is inappropriate to model the trial effects as random effects. Thus the trial effects will be considered as fixed effects in the IPD meta-mediation analysis model. In addition, since there may be further unexplained differences in trial populations, it is sensible to allow for residual heterogeneity across trials. Equivalently, in the IPD meta-mediation analysis model, the variance of the model residuals will be estimated within each of the contributing trials.

Secondly, the differences in interventions and trial design need to be reflected in the IPD meta-mediation analysis model. I will discuss several example scenarios in detail as follows:
Interventions: The SPOKES intervention is a combination of IY and literacy programmes (COMBI), the CPT intervention is an IY only intervention, and HCA includes IY only, literacy only, and COMBI interventions. To take these intervention differences into account, the IY, literacy and COMBI interventions should be represented by different variables in the analysis model. Ideally, the differences in control conditions (if they exist) should also be considered in the analysis model. However, the control conditions cannot be distinguished from the difference of populations (trial effects) in our case, as unique controls were assigned to each of the trials. Since the control conditions are very similar across the three parenting trials, I

assume that the control groups are the same for all three contributing trials in the analysis model.

Changes in randomisation ratios: As mentioned in Section 5.1.2, the three parenting trials, SPOKES, CPT and HCA, employ different trial designs. In HCA, although participants within each recruitment cohort were randomised, the possible treatment groups into which they can be randomised are different in different cohorts, i.e. participants were randomised to intervention groups 1 and 2 in the first cohort and were randomised to groups 3 and 4 in the second cohort. Under this design, recruitment cohort (randomisation ratio) confounds the effect of intervention groups 1 or 2 versus intervention groups 3 or 4. Thus, variables indicating groups with different randomization ratios need to be included in the IPD model for the HCA trial to adjust for such confounding. CPT employed cluster randomisation with an overall treated vs. control randomisation ratio of 2:1 and participants in SPOKES were randomised to treatment and control groups with each school stratum with an overall ratio of 1:1. There is no indication of changes in randomisation ratios. Thus, the randomisation ratios are considered to be constant for CPT and SPOKES.

Randomisation stratifier: Stratified randomisation is commonly used to achieve the balance of participants with certain characteristics in the treated and control groups. If the stratifier is thought to have an impact on the effect of treatment, it should be adjusted in the analysis model as fixed or random effects. The statistical methods of adjusting stratifiers have been discussed by Kahan and Morris' literature (Kahan and Morris, 2013). SPOKES randomisation was performed within each school stratum, considering the number of strata (ten strata) and for the generalization of statistical inference, SPOKES strata are adjusted in the model as random effects.

Clusters: The design-related hierarchical data structure of each contributing trial (see Table 2-3, Table 2-5 and Table 2-7 in Chapter 2) should be considered and accounted for, such as allowing for random intercepts at the level of the clusters in the cluster-randomised CPT, at the level of school in SPOKES, and at the level of recruitment cohorts in HCA. Very small intra-cluster correlations (ICCs) indicate that the effects of a cluster factor can be ignored; otherwise they should be included in the IPD model. Therapy groups are the lower level clusters that are applicable in the treatment arm only. In the IPD meta-mediation analysis of

this chapter, the therapy groups in the active treatment arm are included in the model as IVs with specific IV assumptions, as discussed in Chapter 5.

Thirdly, the full IPD meta-mediation model should allow between-trial heterogeneity in the effects of interest and the nuisance parameters. To model the variability of the effects across trials, a different parameter represents the effect of a given explanatory variable on a given response in each trial. In other words, if there are m trials and p parameters describing effects that are assumed to vary across trials, then $m \times p$ variables (and associated parameters) will be included in the IPD model to model trial heterogeneity in these effects.

To illustrate the implementation of the principles for setting up the full IPD model, I use the example of an IV mediation model conditioning on measured confounding variables $\mathbf{X}$ as follows:
For the purpose of comparison with subsequent equation, firstly I list the population model of a general IV mediation analysis for single trial:

$$Y = i_Y + \mathbf{X}\boldsymbol{\delta} + \gamma R + \beta M + \boldsymbol{\Psi u} + \varepsilon_Y \qquad \text{Equation 5-6}$$

$$M = i_M + \mathbf{X}\boldsymbol{\theta} + \alpha R + \mathbf{Z}\boldsymbol{\zeta} + \boldsymbol{\Psi w} + \varepsilon_M \qquad \text{Equation 5-7}$$

where R denotes randomly allocated treatment, M denotes the mediator, Y denotes the clinical outcome and $\mathbf{X}$ represents a set of r confounding variables, so that $\mathbf{X} = (X_1, X_2, \dots, X_r)$. The coefficient γ represents the strength of the relationship between R and Y when holding M at a fixed level; β represents the strength of the relation between M and Y within fixed levels of R; and α is the coefficient representing the strength of the relationship between R and M. The vectors $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ present a set of coefficients indicating the strength of the relationship between $\mathbf{X}$ and Y, and $\mathbf{X}$ and M respectively. The instrumental variables $\mathbf{Z} = (Z_1, Z_2, \dots, Z_k)$ and their corresponding coefficients are indicated by vector $\boldsymbol{\zeta}$**.** The intercepts in each equation are i_Y and i_M respectively; $\boldsymbol{\Psi}$ represents the design feature of the trials, such as cluster structure in the random effect; $\boldsymbol{u}$ and $\boldsymbol{w}$ represent the random effects; ε_Y and ε_M represent the residual errors.

Now, considering the case of three contributing trials $(\mathrm{A}, \mathrm{B}, \text{and } \mathrm{C})$, the full model for the IPD meta-mediation analysis is constructed as below:

$$Y = i_Y + \mathbf{T\tau} + \mathbf{X\delta} + \mathbf{R\gamma} + \mathbf{M\beta} + \mathbf{\Psi u} + \boldsymbol{\varepsilon}_{\mathbf{Y}} \quad \text{Equation 5-8}$$

$$M = i_M + \mathbf{T\rho} + \mathbf{X\theta} + \mathbf{R\alpha} + \mathbf{Z\zeta} + \mathbf{\Psi w} + \boldsymbol{\varepsilon}_{\mathbf{M}} \quad \text{Equation 5-9}$$

Both mediator and outcome models include the trial variables $\mathbf{T} = (T_B, T_C)$, in which T_B and T_C denote two binary variables indicating the trial membership. The coefficients in $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ represent the effects of trial on the outcome Y and on the mediator M respectively, where $\boldsymbol{\tau} = (\tau_B, \tau_C)^{\mathrm{T}}$ and $\boldsymbol{\rho} = (\rho_B, \rho_C)^{\mathrm{T}}$. (Here superscript T indicates the matrix transposition). $\mathbf{X}$ represents three sets of r confounding variables for three contributing trials, so that $\mathbf{X} = (X_1^A, \dots, X_r^A, X_1^B, \dots, X_r^B, X_1^C, \dots, X_r^C)$. All X^A are zeros for participants of trials B and C, all X^B are zeros for participants of trials A and C, and all X^C are zeros for participants of trials A and B. Their corresponding coefficients in the outcome and the mediator models are presented by $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ respectively, where $\boldsymbol{\delta} = (\delta_1^A, \dots, \delta_r^A, \delta_1^B, \dots, \delta_r^B, \delta_1^C, \dots, \delta_r^C)^{\mathrm{T}}$ and $\boldsymbol{\theta} = (\theta_1^A, \dots, \theta_r^A, \theta_1^B, \dots, \theta_r^B, \theta_1^C, \dots, \theta_r^C)^{\mathrm{T}}$. $\mathbf{R}$ is the randomised treatment group indicator for each trial and allowing for multiple treatment groups within a given trial, so that $\mathbf{R} = (R_1^A, \dots, R_{j-1}^A, R_1^B, \dots, R_{h-1}^B, R_1^C, \dots, X_{l-1}^C)$ where j, h, l represent the number of treatment groups for each trial. The coefficients of $\mathbf{R}$ are $\boldsymbol{\gamma} = \left(\gamma_1^A, \dots, \gamma_{j-1}^A, \gamma_1^B, \dots, \gamma_{h-1}^B, \gamma_1^C, \dots, \gamma_{l-1}^C\right)^{\mathrm{T}}$ in the outcome models and $\boldsymbol{\alpha} = \left(\alpha_1^A, \dots, \alpha_{j-1}^A, \alpha_1^B, \dots, \alpha_{h-1}^B, \alpha_1^C, \dots, \alpha_{l-1}^C\right)^{\mathrm{T}}$ in the mediator model. $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are the DE and ETM respectively for each treatment and each trial. The coefficients of mediator $\mathbf{M} = (M^A, M^B, M^C)$ in the outcome model are presented by $\boldsymbol{\beta} = (\beta^A, \beta^B, \beta^C)^{\mathrm{T}}$ and $\boldsymbol{\beta}$ are the EMO for each trial. I have assumed that there is no interaction between treatment and EMO. The instrumental variables for the three trials are represented by $\mathbf{Z} = (Z_1^A, \dots, Z_k^A, Z_1^B, \dots, Z_p^B, Z_1^C, \dots, Z_q^C)$ and the corresponding coefficients are $\boldsymbol{\zeta} = \left(\zeta_1^A, \dots, \zeta_k^A, \zeta_1^B, \dots, \zeta_p^B, \zeta_1^C, \dots, \zeta_q^C\right)^{\mathrm{T}}$, where $k, p,$ and q represent the number of instruments for each trial. Similar to $\mathbf{X}$, R^A, M^A and Z^A are zeros for participants of trials B and C; R^B, M^B and Z^B are zeros for participants of trials A and C; and R^C, M^C and Z^C are zeros for participants of trials A and B. i_Y and i_M are the intercepts of the outcome and mediator models respectively. $\boldsymbol{\Psi} = (\mathbf{\Psi^A}, \mathbf{\Psi^B}, \mathbf{\Psi^C})$ represents the design feature of the trials, such as cluster structure in the random effect; $\boldsymbol{u} = (\boldsymbol{u^A}, \boldsymbol{u^B}, \boldsymbol{u^C})$ and $\boldsymbol{w} = (\boldsymbol{w^A}, \boldsymbol{w^B}, \boldsymbol{w^C})$ represent the random effects. $\boldsymbol{\varepsilon}_{\mathbf{Y}} = (\varepsilon_Y^A, \varepsilon_Y^B, \varepsilon_Y^C)$ and $\boldsymbol{\varepsilon}_{\mathbf{M}} = (\varepsilon_M^A, \varepsilon_M^B, \varepsilon_M^C)$ are the residuals of the outcome model and mediator model respectively.

Briefly, compared to the mediation model for a single trial (Equation 5-6 and Equation 5-7), the three-trial IPD meta-mediation model (Equation 5-8 and Equation 5-9) includes additional trial membership indicators to model the outcome differences between trials in the control condition. It takes one of the trials (Trial A) as the reference trial and the effects of trial B and trial C are modelled as two fixed effects. Since each RCT has one control group, potential differences between the control conditions used in different trials are in fact embedded in the effects of the trials. Thus, when interpreting trial effects, we need to be aware that these comprise effects of varying control conditions. Additionally, in the full IPD meta-mediation model, the treatment, mediator and confounders are parameterised for each trial separately, allowing for between-trial heterogeneity of these effects. For example, if we were considering p observed confounding variables, then we would need to include $3 \times p$ confounding effect parameters in the three-trial full IPD meta-mediation model. Finally, differences in trial design features are modelled via adding trial-specific covariates and constructing different random effects and residuals for each trial.

5.3.1.2 Simplifying the IPD meta-mediation model using a novel bootstrap test

To systematically select an IPD meta-mediation analysis model, the key is the model simplification based on the results of testing the between-trial heterogeneity of both the parameters of interest and the nuisance parameters. Failure to detect significance of the between-trials heterogeneity will lead us to assume constant effect across different trials. The benefits of empirically testing effect heterogeneity and selecting a model for IPD meta-mediation analysis based on the test results are: (i) provision of insight into which effects vary across trials and associated mechanistic interpretation, and (ii) potential gains in precision for estimating respective effects.

To introduce the systematic approach to selecting the final IPD meta-mediation model, this section will be based on the full IPD meta-mediation model (Equation 5-8 and Equation 5-9) of the previous section. The full model will then be simplified via testing the between-trial heterogeneity of the effects of a set of variables (the intervention, the mediator and the confounders) in both outcome model and mediator model. A novel non-parametric bootstrap approach based on confidence ellipsoids is developed in this section for carrying out such tests of composite interaction null hypotheses.

Taking the numeric mediator $\mathbf{M}$ in the outcome model as an example, in the IPD full model there are in fact three variables, M^A, M^B, and M^C, that represent the same mediator for trials A, B and C respectively. Their corresponding coefficients are β^A, β^B and β^C in the outcome model. I am interested to know if the effects of the mediator on the outcome are constant across trials. I would therefore like to test the following null hypothesis:

$$H_0: \beta^A = \beta^B = \beta^C \qquad \text{Equation 5-10}$$

The alternative hypothesis is

$$H_1: \beta^A \neq \beta^B \text{ or } \beta^A \neq \beta^C \qquad \text{Equation 5-11}$$

In difference terms, we have the following composite hypotheses:

$$H_0: \beta^A - \beta^B = 0 \text{ and } \beta^A - \beta^C = 0 \qquad \text{Equation 5-12}$$

$$H_1: \beta^A - \beta^B \neq 0 \text{ or } \beta^A - \beta^C \neq 0 \qquad \text{Equation 5-13}$$

In vector notation:

$$H_0: \begin{pmatrix} \beta^A - \beta^B \\ \beta^A - \beta^C \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{Equation 5-14}$$

$$H_1: \begin{pmatrix} \beta^A - \beta^B \\ \beta^A - \beta^C \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{Equation 5-15}$$

The estimator of $\phi = \begin{pmatrix} \beta^A - \beta^B \\ \beta^A - \beta^C \end{pmatrix}$ is $\hat{\phi} = \begin{pmatrix} \hat{\beta}^A - \hat{\beta}^B \\ \hat{\beta}^A - \hat{\beta}^C \end{pmatrix} = \begin{pmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \end{pmatrix}$. If there are B bootstrap samples simulated from the observed data, then the vector $\hat{\phi}^b = \left(\hat{\phi}_1^b, \hat{\phi}_2^b\right)^T$ generated for each of B bootstrap samples is equivalent to an observation point $x = (x_1, x_2)^T$ of a group of B observations with mean $\mu = (\mu_1, \mu_2)^T$ with the variance-covariance matrix of the data $\mathbf{V}$ as introduced in Section 5.2.4. The centre of mass μ of this bootstrap distribution is provided by the original data estimator $\hat{\phi}^o = \left(\hat{\phi}_1^o, \hat{\phi}_2^o\right)^T$ and the variance-covariance matrix is estimated by the bootstrap samples' variance-covariance matrix $\mathbf{S}$. Based on this information, we can calculate the Mahalanobis distances (MDs) from the bootstrap points or the zero point $(0,0)^T$ to the reference central point $\hat{\phi}^o = \left(\hat{\phi}_1^o, \hat{\phi}_2^o\right)^T$ using Equation 5-16.

$$D_M(\hat{\phi}) = \sqrt{(\hat{\phi} - \hat{\phi}^o)^T \mathbf{S}^{-1} (\hat{\phi} - \hat{\phi}^o)} \qquad \text{Equation 5-16}$$

Following Equation 5-4, the MD can be used as the pivot statistic for testing the null hypothesis $\phi_1 = \phi_2 = 0$. The p value can be calculated as

$$p = \Pr(D_M^{*2} \geq d_0^2 | \hat{F}) \qquad \text{Equation 5-17}$$

where D_M^* is the Mahalanobis distance calculated for each of B bootstrap samples, d_0 is the Mahalanobis distance from the zero point $(0,0)^T$ to the reference (centre of mass) point $\left(\hat{\phi}_1^o, \hat{\phi}_2^o\right)^T$, and $\hat{F}$ is the bootstrap sampling distribution of the parameter estimate $\hat{\phi}$. Intuitively, we are testing whether the zero point $(0,0)^T$ is in the confidence ellipsoid formed by the bootstrap estimates points. If we had the bootstrap estimated p quantiles d_p^* Mahalanobis distance from the centre of mass, then the $100 \times (1-\alpha)\%$ joint confidence region for ϕ would be all ϕ for which $D_M \leq d_{1-\alpha}^*$. If the d_0^2 is greater than or equal to $(d_{1-\alpha}^*)^2$, the null hypothesis is rejected. On the contrary, if the d_0^2 is less than $(d_{1-\alpha}^*)^2$, the null hypothesis $\beta^A = \beta^B = \beta^C$ is not rejected. This means that the effect of the mediator on the outcome does not differ significantly between the trials at the $\alpha\%$ level. Thus, a more parsimonious model description is given by a constant mediator effect across trials and the M^A, M^B and M^C can be combined as one variable M with only one coefficient scalar.

Assessing trial-heterogeneity of the causal effects of interest is part of the process of IPD meta-mediation analysis. It can provide answers to a list of questions:

- Do observed confounders operate in the same way across trials? We would expect this if their effect represented a mechanism that operates in a population. We can assess this empirically by testing whether both the effect of the confounder on the mediator ($\theta^A = \theta^B = \theta^C$) and its effects on the outcome are constant across trials ($\delta^A = \delta^B = \delta^C$). If the confounder effects are found not to vary, then the associated parameters can be restricted to be constant across trials in the IPD model to improve the power of the mediation analysis.
- Does the effect of the intervention(s) on the mediator or the direct effect on the outcome vary across trials? Depending on the trial variation in the implementation of a complex intervention (e.g. in delivery by therapists with different levels of skill), we might expect to see differences here even for the same target population. We can assess this by testing $\alpha^A = \alpha^B = \alpha^C$ or $\gamma^A = \gamma^B = \gamma^C$ respectively. If no between-trial heterogeneity is found, the parameters of different trials can again be combined as one

parameter, i.e. a constant intervention effect on the mediator across trials (α) and a constant intervention effect on the outcome (γ) will be estimated in the IPD model.

- Does the mediator affect the outcome in the same way across trials? Again, we would expect this if the putative mediator represented a mechanism that operates in a population. Testing the heterogeneity of the effect of the mediator on the outcome (null hypothesis: $\beta^A = \beta^B = \beta^C$) can provide further empirical evidence of the mechanistic qualities of the putative mediator and, if the mediator effect can be assumed to be constant, can provide improved power for mediation analysis.

A backward selection approach is applied to assess the trial heterogeneity of the causal effects of interest. It starts from the full IPD meta-mediation analysis model (Equation 5-8 and Equation 5-9), which allows all the causal effects of interest to vary with trials. The sequence of the tests of this backward selection approach is described as follows: Firstly, I test the trial heterogeneity of a list of confounders included in the full model. For each confounder X_i, I test the hypothesis $\theta_i^A = \theta_i^B = \theta_i^C$ (constant confounding effect on the mediator) and $\delta_i^A = \delta_i^B = \delta_i^C$ (constant confounding effect on the outcome) assuming the existence of trial heterogeneity for the rest of the causal effects of interest (including the other confounders). If neither $\theta_i^A = \theta_i^B = \theta_i^C$ nor $\delta_i^A = \delta_i^B = \delta_i^C$ are rejected, constant confounding effect of X_i across trials is supported. Then the IPD model can be simplified based on the results of trial heterogeneity testing of each confounder. Secondly, I test the trial heterogeneity of the ETM ($\alpha^A = \alpha^B = \alpha^C$) and DE ($\gamma^A = \gamma^B = \gamma^C$) respectively using the confounder-simplified IPD model, assuming the existence of trial heterogeneity for the rest of the causal effects of interest. If the trial heterogeneity of ETM and DE are rejected, the IPD model can be further simplified, i.e. using one parameter (α) to model constant ETM across trials and one parameter (γ) to model constant DE across trials. Finally, I test the trial heterogeneity of the EMO ($\beta^A = \beta^B = \beta^C$) based on the further simplified IPD model obtained from the previous step. The final simplified IPD model will be constructed based on the result of this test.

Simply put, assessing trial heterogeneity of the effects provides an insight into effect-trial interaction. Consequently, it contributes to the construction of the IPD meta-mediation model. Failure to detect effect heterogeneity can be taken as empirical support for combining effects across trials (assuming a constant effect across trials). This leads to a

simplified IPD model and a potential improvement in precision for mediation analysis. Since the constant effect assumption is based on the non-significant results of the trial-heterogeneity tests, the power of the test has impact on the results. This means that the non-significant trial-heterogeneity may be due to lack of power. A liberal significance level (e.g. 10%) may reduce the harm caused by the low power of the test. On the other hand, multiple tests are required to construct the final IPD meta-mediation model, which leads to the issue of multiplicity (when testing multiple times, the probability of obtaining significant results by chance increases). From that point of view, more conservative significant level is required to adjust the type I errors. Thus, a significant level of 5% is selected here as a sensible compromise between lack of power and multiplicity. In the following section, I propose to combine IPD meta-mediation analysis with the IV-MI-BT approach to draw statistical inferences and account for missing values.

5.3.2 Extension of the IV-MI-BT approach to IPD meta-mediation analysis

In this section, I will introduce a strategy that utilises the non-parametric bootstrapping approach to draw statistical inferences in the IPD meta-mediation analysis. It includes making decisions regarding the existence of trial heterogeneity in effects and providing estimation (including standard errors and confidence intervals) of the effects of interest. In addition, the principles of using MI to handle missing data in the IPD meta-mediation analysis will be discussed. In fact, combining IV-MI-BT method with IPD meta-mediation analysis can be viewed as an extension of the IV-MI-BT method to multiple trials. The implementation of the combined IV-MI-BT and IPD meta-mediation analysis approach will also be illustrated in this section.

As discussed in Chapters 3 and 4, statistical inferences for causal mediation parameters are generated using non-parametric bootstrap resampling methods that mimic the single trial data generating process. For an IPD meta-mediation analysis, the question is how to resample data from different trials targeting different populations. Since the contributing trials were run independently, to mimic the data generating process of these trials, I propose to perform a bootstrap resampling procedure within each trial. Following the same bootstrapping principles as introduced in the previous chapters, the resampling approach of each individual trial should reflect the design of that trial (e.g. bootstrapping for hierarchically structured data in Chapter 3). Bootstrapping within contributing trials

facilitates the analysis accounting for differences between trial designs. Section 5.4.3 demonstrates the details of applying different bootstrapping strategies for different trial designs using an example of three parenting programme RCTs for the IPD mediation analysis. The approach will generate B bootstrap samples for each trial and the bootstrap samples of the contributing trials will be combined together to form the final bootstrap samples for drawing statistical inferences. For instance, if three trials A, B and C are included in the IPD analysis, the first bootstrap samples $\mathrm{B}_1^{\mathrm{A}}$, $\mathrm{B}_1^{\mathrm{B}}$ and $\mathrm{B}_1^{\mathrm{C}}$ of the three trial trials are combined to form the first bootstrap sample B_1 for the IPD analysis, $\mathrm{B}_2^{\mathrm{A}}$, $\mathrm{B}_2^{\mathrm{B}}$ and $\mathrm{B}_3^{\mathrm{C}}$ are put together to form B_2, and so forth. At the end, B combined bootstrap samples for IPD analysis are produced.

In line with the IV-MI-BT approach described in the last chapter, MI is applied to the original data sample and to each bootstrap replicate for constructing estimators in the presence of missing values. Specifically for IPD meta-mediation analysis, MI is performed within each trial separately in order to preserve any between-trial variability in parameters. Again, the specific MI procedure applied here is MICE and the procedure in the IPD analysis obeys the same rules as the single trial missing data MI introduced in Chapters 3 and 4. Application of MICE in the IPD mediation analysis using three parenting trials is described in Section 5.4.3.

The implementation of the IV-MI-BT IPD meta-mediation analysis approach consists of five steps: preparing the pooled data for IV-MI-BT IPD meta-mediation analysis; setting up a full IPD meta-mediation model; generation of bootstrap re-samples for pooled data; IPD model simplification; and IPD meta-mediation analysis based on the final model. The details of these five steps will be explained and demonstrated in the following paragraphs.

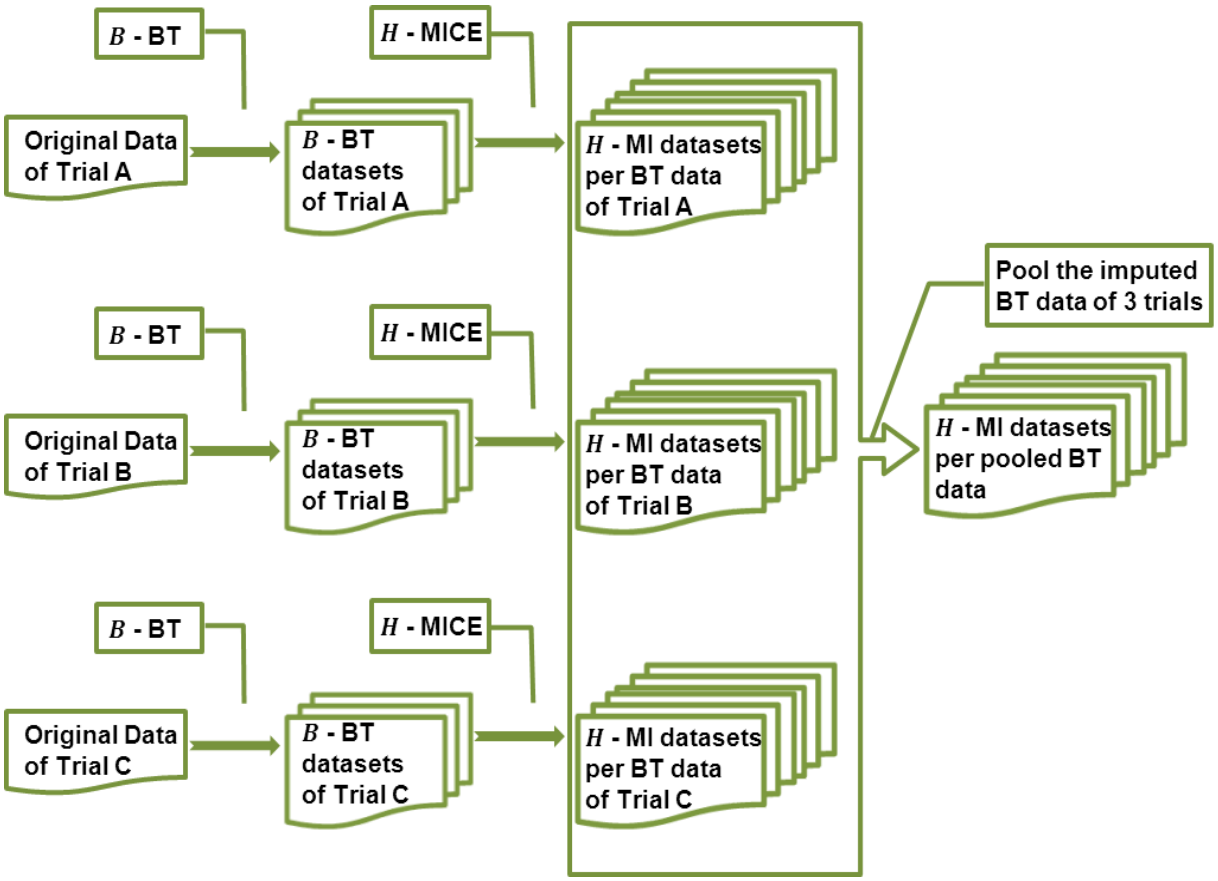
Step 1 – Preparing the pooled data for IV-MI-BT IPD meta-mediation analysis. IPD meta-mediation analysis includes individual level data from multiple contributing studies that investigated similar interventions and employed compatible measures of the mediators and the outcomes. However, the methods and applications of the measurements may not be identical across trials so that harmonisation of the measurements is required. Details of measurement harmonisation have been described in Chapter 2 Section 2.4, using the example of three parenting trials. In the pooled dataset, additional variables need to be constructed for indexing different trials (trial identifiers) and the coding of the interventions

should be consistent across trials for indicating the different/same inventions. As recommended in Chapter 4, for interaction instruments (interactions between treatment groups R and baseline covariates Z) in the IV mediation analysis, an orthogonalisation needs to be performed to ensure that the interpretation of the regression coefficient (α) of the intervention effect in the mediator model remains unchanged when adding the IVs into the model. For pooled data, this IV orthogonalisation is completed using the original data within each trial in the data preparation stage.

Step 2 – Setting up a full IPD meta-mediation model. The full IPD meta-mediation model, as specified in Equation 5-8 and Equation 5-9, includes the trial indicator ($\mathbf{T}$), intervention indicator ($\mathbf{R}$), outcome ($\mathbf{Y}$), putative mediator ($\mathbf{M}$), observed confounders ($\mathbf{X}$), instrumental variables ($\mathbf{Z}$) and variables reflecting the design feature of the trials ($\mathbf{\Psi}$). As the full IPD meta-mediation model allows trial heterogeneity of all the effects of interests, parameterisation of the confounders, putative mediators, IVs and trial design are trial specific. The principles of setting up a full IPD meta-mediation model have been discussed in Section 5.3.1.1.

STEP 3 – Generation of bootstrap re-samples for pooled data. Using an example of three contributing trials (represented by A, B and C), Figure 5-1 demonstrates the procedure of generating data for IV-MI-BT-IPD meta-mediation analysis. As mentioned in Section 5.3.2, bootstrap samples with MI will be used in the IPD mediation analysis for drawing statistical inferences. Basically, bootstrap resampling and Multiple Imputation are conducted within each trial separately following the single trial BT and MI procedure as shown in Section 3.3.2.1 steps 1 and 2. For each trial, B bootstrap samples are generated from the original data, usually with missing values, and then MICE are applied to each bootstrap sample, forming H imputed datasets for each bootstrap sample: that is, $H \times B$ imputed bootstrap samples in total. In this three-trial example, we then get $H \times B$ imputed bootstrap samples for trial A, trial B and trial C respectively. Finally, the bootstrap re-samples for pooled data are generated by pooling all imputed bootstrap samples of the three trials together.

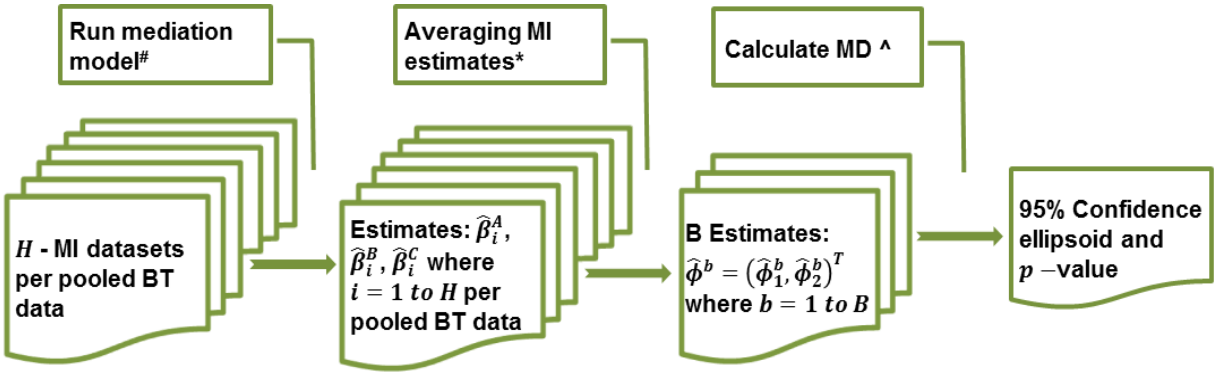
Figure 5-1 MI-BT data generating procedure for three contributing trials



Step 4 – IPD model simplification via testing trial heterogeneity of effects of the mediators (β), the interventions (α, γ) and the confounders (θ, δ). The proposed novel non-parametric bootstrap approach in Section 5.3.1.2 for testing multiple parameters is applied for the model simplification. The backward selection approach is applied to assess the trial heterogeneity of the causal effects of interest. The trial heterogeneity of confounders is tested first, followed by the tests of trial-heterogeneity of the interventions, and the trial heterogeneity of the effects of the mediators is tested at the end. Details of the hypothesis and the assumptions of the tests have been discussed in Section 5.3.1.2. The final IPD meta-mediation model is obtained at the end of this simplification procedure.

Taking the effect of the mediator on the outcome β in the case of three contributing trials as an example, Figure 5-2 is generated to illustrate the procedure for testing the effect heterogeneity over trials.

Figure 5-2 MI-BT procedure for testing the effect heterogeneity across trials



The mediation model allows the trial heterogeneity of the EMO $(\beta^A, \beta^B, \beta^C)$

* $\hat{\phi}_1 = 1/h \sum_1^h (\hat{\beta}_i^B - \hat{\beta}_i^A)$ and $\hat{\phi}_2 = 1/h \sum_1^h (\hat{\beta}_i^C - \hat{\beta}_i^B)$

^ Mahalanobis Distance (MD): $D_M(\hat{\phi}^b) = \sqrt{(\hat{\phi}^b - \hat{\phi}^0)^T S^{-1} (\hat{\phi}^b - \hat{\phi}^0)}$, where the centre of mass of this bootstrap distribution is provides by the original data estimator $\hat{\phi}^o = (\hat{\phi}_1^o, \hat{\phi}_2^o)^T$ and S is the bootstrap sample variance-covariance matrix, $\hat{\phi}^0$

Firstly, the MI-BT imputed re-samples were used to run the mediation model allowing trial heterogeneity of the effect of interest (β in this example). For H impuations per BT data, the model estimated H estimates of $\hat{\beta}^A, \hat{\beta}^B, and\ \hat{\beta}^C$. After taking the average of H estimates for testing the composite hypothesis ($\hat{\beta}^A = \hat{\beta}^B = \hat{\beta}^C$), I get $\hat{\phi}_1 = 1/h \sum_1^h (\hat{\beta}_i^B - \hat{\beta}_i^A)$ and $\hat{\phi}_2 = 1/h \sum_1^h (\hat{\beta}_i^C - \hat{\beta}_i^B)$ for a bootstrap sample. In total, I have B bootstrap samples, and for each BT sample I can calculate $\hat{\phi}^b = (\hat{\phi}_1^b, \hat{\phi}_2^b)^T$ where $b = 1\ to\ B$. Then I can use $D_M(\hat{\phi}^b) = \sqrt{(\hat{\phi}^b - \hat{\phi}^0)^T S^{-1} (\hat{\phi}^b - \hat{\phi}^0)}$ to calculate the Mahalanobis distance, where the centre of mass of this bootstrap distribution is provided by the original data estimator $\hat{\phi}^o = (\hat{\phi}_1^o, \hat{\phi}_2^o)^T$ and S is the bootstrap sample variance-covariance matrix. Finally, I can generate the 95% confidence ellipsoid and the p-value based on the empirical distribution of $D_M(\hat{\phi}^b)$. If the 95% confidence ellipsoid includes zero point, the null-hypothesis $\beta^A = \beta^B = \beta^C$ is not rejected, and the most efficient estimator of the effect of the mediator on the outcome will be the synthesised single estimator β.

Step 5 – Analysing the final IPD meta-mediation model using the IV-MI-BT approach. This step provides the synthesised point estimates and the corresponding confidence intervals for the effects of interest based on fitting the simplified IPD mediation model. The procedure of the IV-MI-BT IPD meta-mediation analysis is the same as the one demonstrated in Figure 3-2 of Chapter 3 apart from two aspects: 1) the MI-BT data generating procedure is the procedure introduced in Step 3, and 2) the mediation analysis model is the final simplified IPD meta-mediation analysis model including instrumental variables. Importantly, the final non-parametric IPD meta-mediation analysis accounts for trial variability and is valid in the presence of missing values provided that the data generating process is MAR.

5.4 Application of the IV-MI-BT approach for IPD meta-mediation analysis to the SPOKES, CPT and HCA studies

In this section, meta-mediation analysis of IPD from three parenting trials - SPOKES (Scott et al., 2010b) , CPT (Scott et al., 2001b) and HCA (Scott et al., 2012a) - will be carried out using the IV-MI-BT approach. The participants, design, intervention and measurements of the three studies have been introduced in Chapter 2 of this thesis. The current example IPD meta-mediation analysis investigates the pre-specified putative mediators (expressed criticism and expressed warmth as identified in Chapter 3 on the basis of the SPOKES data alone) and also assumes that variables acting as observed confounders or instrumental variables have been identified (based on the SPOKES investigations in Chapters 3 and 4 respectively).

5.4.1 Preparation of the pooled data set comprising three parenting studies

The pooled data set for the SPOKES, CPT and HCA trials consists of the following individual level variables (see Chapter 2):

1) Trial identifiers: Two dummy variables were created as trial identifiers, taking CPT as the reference trial.
2) Treatment variables: IY only, literacy only, IY and literacy combined.
3) Clinical outcome variables: Parent accounted child conduct problem.
4) Putative mediator variables: parental expressed criticism and parental expressed warmth.

5) Observed baseline confounders: child's gender, child's reading ability, parent's education, parent's depression, lone parent, the putative mediator measured at baseline, child outcome measured at baseline.
6) Instrumental variables for expressed criticism: intervention × baseline parental depression interaction, intervention × baseline parental education interaction, and therapy groups in the active intervention arm.
7) Instrumental variables for expressed warmth: number of sessions attended (%) in the active intervention arm.
8) Variables representing design features: SPOKES: school-year strata; CPT: randomisation clusters; HCA: recruitment cohorts.
9) Auxiliary variables: variables measured at baseline but not selected as confounders, parenting practice measurements other than the putative mediators, child behaviour other than the primary clinical outcome.

Following the framework of the IPD meta-mediation analysis using the IV-MI-BT approach proposed in Section 5.3.2, the first step is preparing the IV-MI-BT data. This step includes the orthogonalisation of the IVs (interaction terms) for SPOKES, CPT and HCA separately. Further details of the interaction term orthogonalisation process can be found in Section 4.2 of Chapter 4.

5.4.2 Setting up a full IPD meta-mediation model

The systematic approach to constructing the IPD meta-mediation model begins with setting up the full IPD meta-mediation model accounting for the trial main effects, the between-trial heterogeneity of the effects of interest and the variability of trial designs following the principles discussed in Section 5.3.1.1. In this project, the effect of three parenting studies – CPT, SPOKES and HCA – will be included in both the mediation (variables in Section 5.4.1 list 4.) and the outcome model (variable in Section 5.4.1 list 3.) as fixed effects ($\mathbf{T}$ in Equation 5-8 and Equation 5-9) and by allowing the residual variance ($\boldsymbol{\varepsilon}_{\mathbf{Y}}$ and $\boldsymbol{\varepsilon}_{\mathbf{M}}$) to vary with trial. The set of seven baseline confounders are also included in both the mediator and the outcome models (variables in Section 5.4.1 list 5.). Assuming the existence of between-trial variability in confounding effects, a unique variable is given to each confounder in each trial ($\mathbf{X}$ in Equation 5-8 and Equation 5-9). As we know, the three contributing trials delivered different active interventions. Specifically, these were IY only intervention in CPT, IY and literacy combined intervention in SPOKES, and three active interactions in HCA: IY only,

literacy only, and IY and literacy combined. Thus, five binary variables are used to present the effects of different active interventions in different trials ($\mathbf{R}$ in Equation 5-8 and Equation 5-9). The control groups of the three trials are considered as the reference group. Note that for the current three trials, the effects of different control conditions (telephone helpline control in SPOKES; waiting list control in CPT; service as usual/'signposting' control in HCA) cannot be distinguished from trial effects, as each trial has only one control group. Since the control conditions are very similar across the three parenting trials, I assume that the control groups are the same for all three contributing trials in the analysis model. In each trial, the IVs for the putative mediator expressed criticism are the intervention × baseline parental depression interaction, the intervention × baseline parental education interaction, and the therapy groups in the active intervention arm; the IVs for the putative mediator expressed warmth are the number of sessions attended (%) in the active intervention arm. In the full IPD meta-mediation model, the mediator $\mathbf{M}$ and the respective instruments $\mathbf{Z}$ will be represented by a set of trial-specific variables.

Finally, variables need to be included to account for design features of the three trials: For HCA specifically, a set of binary variables indicating the change of the randomisation ratios across recruitment cohorts are included in both the mediator and the outcome models to adjust for potential confounding caused by the different cohorts having different sets of treatment groups to which the participants can be randomised. It is considered that the randomisation ratios are consistent in SPOKES and CPT. Stratified randomisation was applied in SPOKES. For the generalisation of statistical inference, the school-year stratifiers (ten strata) are included in the analysis model as random effects. For SPOKES, this may overlap with the method of accounting for clustered data as follows. As reviewed in Chapter 2, the three parenting trials have hierarchical data structures. To model the hierarchical data structures in the three trials, random effects are considered for each trial. Theoretically, the highest levels of clustering in each trial should also be reflected by a further random effect varying at that level. As introduced in Chapter 2, the highest level clusters of the three trials are the randomisation clusters in CPT, the school-year strata in SPOKES, and the recruitment cohorts in HCA. However, these variance components might be negligible and in practice very small variance components can lead to convergence problems when running the analysis model. In order to address this practical issue, I calculated the intra-class correlation (ICC) coefficient of the highest-level cluster variable within each trial to judge the size of the

highest-order cluster effects. More specifically, IV mediation models with the highest level clusters as random intercepts were run for each trial and each putative mediator separately. The estimates of the variances of the random intercepts, the variances of the residuals and the resulting ICC coefficients of the highest-level cluster variables are shown in Table 5-1.

Table 5-1 The variances of the random effects and resulting ICCs of the highest-level cluster variables

		Mediator model			Outcome model		
Mediator	Trial	Variance of random intercepts	Variance of residuals	Intra class correlation (ICC)	Variance of random intercepts	Variance of residuals	Intra class correlation (ICC)
Expressed Criticism	CPT	3.79E-24	0.76	5.26E-24	9.44E-18	0.15	6.52E-17
	SPOKES	1.95E-18	0.23	8.23E-18	1.15E-17	0.09	1.16E-16
	HCA	4.53E-23	0.37	1.23E-22	2.83E-24	0.15	1.86E-23
Expressed Warmth	CPT	0.03	0.60	0.05	2.68E-18	0.15	1.74E-17
	SPOKES	0.02	0.38	0.04	5.24E-21	0.10	5.52E-20
	HCA	1.26E-22	0.45	2.77E-22	3.52E-24	0.15	2.24E-23

The ICC coefficients of the highest-level clusters within each trial are trivial in the mediator and the outcome models for both putative mediators. The estimates of the variances of the cluster-effects agreed with the ICC coefficients calculated by fitting mediation models for each trial separately. These results lend empirical support for assuming zero higher-order cluster effects in the analyses of these three data sets. Thus in the full IPD meta-mediation analysis model, no random effects other than the residuals (allowing for between-trial heterogeneity in residual variances) are included. Therapy groups in the active treatment arm (the low-level clusters) are included in the IPD meta-analysis model as IVs.

5.4.3 Generating bootstrap samples of the pooled data set

The multiple-imputed bootstrap samples are generated within each trial separately following the procedure shown in Figure 5-1. As introduced previously, the three parenting trials have hierarchical data structures and their trial designs are slightly different. The bootstrap re-sampling strategy in the presence of clusters is to resample at the highest cluster level. For the cluster randomised trial CPT, this means randomly sampling clusters of participants with replacement. As the randomisation clusters (Chapter 2, Section 2.2.2) define the highest

grouping level, they are the resampling units. This resampling strategy retains the correlation of the individuals within the same cluster or sub-cluster and it also maintains the trial's randomisation ratio in the bootstrap samples. For HCA, this means sampling randomly with replacement at the level of recruitment cohorts (Chapter 2, Section 2.2.3). For SPOKES, this means resampling at the level of school-year strata (Chapter 2, Section 2.1.2). This is the same resampling strategy that has been applied in Chapters 3 and 4.

In order to ensure that the imputed data allows the effect heterogeneity across trials, MI is performed separately for each trial's bootstrap sample. Although the variables for estimating the effects of interest have been collected in all three trials, the different trials had different trial design variables and collected different set of auxiliary variables that can be used to predict the missingness. The imputation models for SPOKES IV-MI-BT mediation analysis have been described in detail in Section 4.4.3 of Chapter 4. The imputation models of CPT and HCA are similar to the imputation models of SPOKES. Compared with the SPOKES imputation models, CPT models excluded the questionnaire measures of parenting practices from the imputation model, as they were not measured in this trial. For the design variables, CPT models included the randomisation clusters in the imputation models. Correspondingly, the HCA trial imputation model included the randomisation recruitment cohorts but not the school-year strata compared with the SPOKES trial.

5.4.4 Selection of the final IPD meta-mediation model

After running the full IPD meta-mediation model using the IV-MI-BT approach, it generates estimates of the effects of interest ($\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}$) and of nuisance parameters ($\hat{\boldsymbol{\tau}}, \widehat{\boldsymbol{\rho}}, \widehat{\boldsymbol{\delta}}, \widehat{\boldsymbol{\theta}}$) from the original sample and the BT samples. Although the estimates of the error variance parameters are also generated from the model, they are not my research interest. As the full model allows all these effects to vary with trial, their estimates from the bootstrap samples can be used to test effect variability over trials (between-trial heterogeneity).

To investigate the between-trial heterogeneity of the effects of interest with the observed confounders and select a final model for integrative analysis, the novel bootstrap test described in Section 5.3.1.2 is applied. Specifically, a set of non-parametric bootstrap tests based on the confidence ellipsoid methods are applied. The following paragraphs displayed the hypotheses to be tested, the order of the tests, and the test results with interpretations.

Firstly, I tested the between-trial heterogeneity of the confounding effects for the purpose of power improvement via combining effects across trials. For each confounder X_i, I test the hypothesis $\theta_i^A = \theta_i^B = \theta_i^C$ (constant confounding effect on the mediator) and $\delta_i^A = \delta_i^B = \delta_i^C$ (constant confounding effect on the outcome) assuming the existence of trial heterogeneity for the rest of the causal effects of interest (including the other confounders). The results of the non-parametric bootstrap tests based on the squared Mahalanobis distance (confidence ellipsoid) are shown in Table 5-2 for testing the trial-heterogeneity of the confounders in the mediation analyses for expressed criticism and expressed warmth respectively.

Table 5-2 Confounding effects between-trial heterogeneity

		Mediator Model			Outcome Model		
Mediators	Confounders	MD value	95% threshold	p-value	MD value	95% threshold	p-value
Expressed warmth	Baseline child behaviour	1.49	6.16	0.48	0.12	0.95	0.2
	Baseline warmth	3.33	6.47	0.2	0.06	1.1	0.35
	Parent depression	9.9	6.71	0.01*	0.1	1.72	0.57
	Child reading ability	1.11	7.27	0.53	0.28	1.93	0.15
	Lone parent	3.73	5.99	0.15	0.24	1.14	0.2
	Child gender	0.05	6.22	0.97	0.14	0.57	0.18
	Parent education	0.45	6.04	0.8	0.1	1.36	0.56
Expressed criticism	Baseline child behaviour	1.31	7.61	0.51	2.15	6.07	0.37
	Baseline criticism	10.44	6.42	0.01*	1.21	6.18	0.54
	Parent depression	6.03	6.41	0.06	1.92	6.41	0.38
	Child reading ability	2.62	6.55	0.25	15.64	6.64	0.01*
	Lone parent	0.38	6.07	0.83	4.51	7.16	0.16
	Child gender	2.86	5.85	0.26	6.34	6.83	0.07
	Parent education	5.46	7.28	0.08	1.54	6.13	0.47

Note: MD represents the squared Mahalanobis distance; p-values < 0.05 are asterisked; the confounders with significant between-trial heterogeneity in either the mediator model or the outcome model are highlighted in grey.

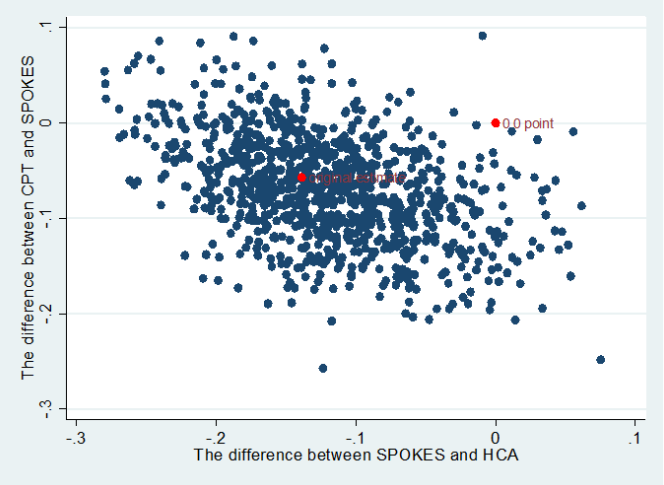
The effects of parent depression on expressed warmth mediator are significantly different at the 5% level over the three parenting trials. In the expressed criticism mediation analysis

model, the effect of baseline expressed criticism on the mediator and the effect of child's reading ability on the child outcome showed significant effects variability at the 5% level over the three parenting trials. The trial heterogeneity of the effect of the confounders might be due to the different target population. No significant differences across trials were found for the other baseline confounders. Therefore, in the warmth mediation analysis, the confounding effects of parental depression will still be modelled by three parameters (one for each trial), as will the confounding effects of baseline criticism and child reading ability in the criticism mediation analysis. On the other hand, the confounding effects will be combined across trials if the baseline confounder shows no significant between-trial heterogeneity. For example, I will use a single parameter, $\theta_{\text{child gender}}$, to represent the (average) effect of child gender on the mediator and a single parameter, $\delta_{\text{child gender}}$, to represent the effect of child gender on the outcome in the combined population targeted by the three trials.

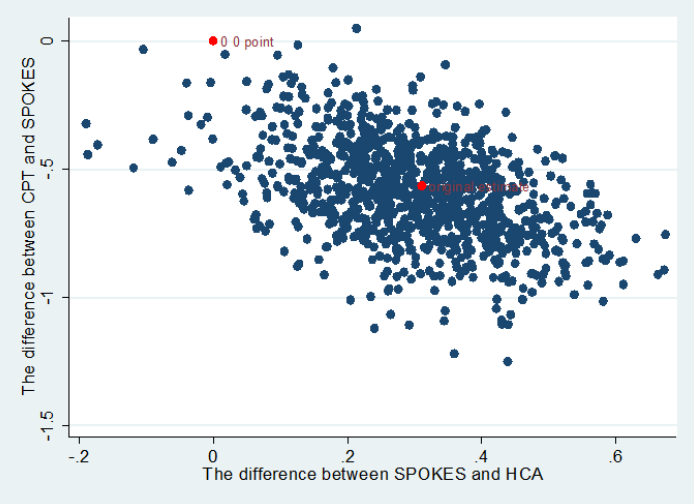
The between-trial differences (CPT vs. SPOKES; SPOKES vs. HCA) in confounder effects are illustrated by three scatter graphs (Figure 5-3) for the effect of parent depression in the warmth mediator model, the effect of baseline criticism in the criticism mediator model, and the effect of child reading ability in the outcome model for expressed criticism respectively. They are the three baseline confounding variables with significant between-trial heterogeneity. The reference central mass points and the zero points are highlighted in red. It is obvious that the zero points are located far from the centre of the scattered points in all three cases. These scatter plots aim to provide a straightforward graphical expression of the significant between-trial heterogeneity of the selected confounders.

Figure 5-3 Scatter graphs illustrating results of bootstrap test for trial heterogeneity in confounder effects

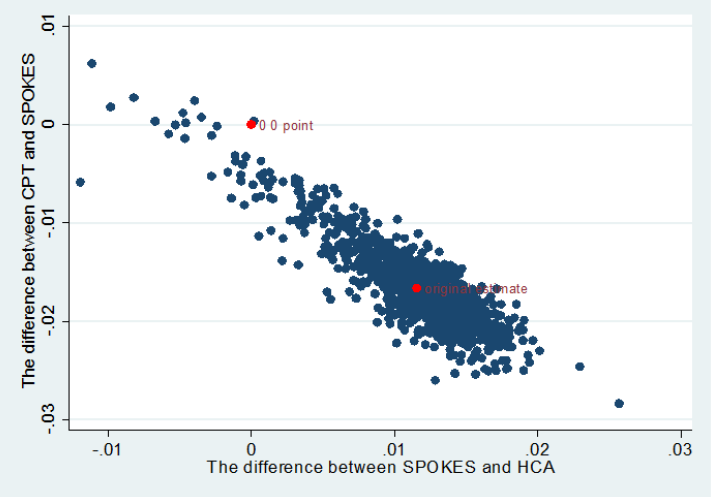
Parental depression in the MEDIATOR model for EE warmth



EE criticism at baseline in the MEDIATOR model for EE criticism



Child reading ability in the OUTCOME model for EE criticism



Secondly, I proceed to test the between-trial heterogeneity of the causal mediation effects of interest ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$) within each intervention. The IY intervention was delivered in the CPT and HCA trials, and the IY and literacy combined intervention was delivered in the SPOKES and HCA trials. Thus, the heterogeneity hypotheses I am going to test are $\alpha_{\mathrm{IY}}^{\mathrm{CPT}} - \alpha_{\mathrm{IY}}^{\mathrm{HCA}} = 0$, $\beta_{\mathrm{IY}}^{\mathrm{CPT}} - \beta_{\mathrm{IY}}^{\mathrm{HCA}} = 0$, $\gamma_{\mathrm{IY}}^{\mathrm{CPT}} - \gamma_{\mathrm{IY}}^{\mathrm{HCA}} = 0$ and $\alpha_{\mathrm{combi}}^{\mathrm{SPOKES}} - \alpha_{\mathrm{combi}}^{\mathrm{HCA}} = 0$, $\beta_{\mathrm{combi}}^{\mathrm{SPOKES}} - \beta_{\mathrm{combi}}^{\mathrm{HCA}} = 0$, $\gamma_{\mathrm{combi}}^{\mathrm{SPOKES}} - \gamma_{\mathrm{combi}}^{\mathrm{HCA}} = 0$. These hypotheses are tested separately, assuming the existence of trial heterogeneity of the other effects of interest. The MI-BT pivot method is applied to test these hypotheses. The results of the tests are listed in Table 5-3.

Table 5-3 Tests for trial-heterogeneity of mediation effects of interest

Putative Mediator	Test Parameter	Estimate	SE	P-value	Percentile 95% BT CI
Warmth	$\alpha_{\mathrm{IY}}^{\mathrm{CPT}} - \alpha_{\mathrm{IY}}^{\mathrm{HCA}}$	-0.35	0.29	0.21	(-0.93, 0.2)
	$\beta_{\mathrm{IY}}^{\mathrm{CPT}} - \beta_{\mathrm{IY}}^{\mathrm{HCA}}$	1.79	9.41	0.37	(-6.09, 12.04)
	$\gamma_{\mathrm{IY}}^{\mathrm{CPT}} - \gamma_{\mathrm{IY}}^{\mathrm{HCA}}$	-0.42	4.7	0.69	(-4.45, 3.18)
	$\alpha_{\mathrm{combi}}^{\mathrm{SPOKES}} - \alpha_{\mathrm{combi}}^{\mathrm{HCA}}$	-0.38	0.33	0.23	(-1.23, 0.08)
	$\beta_{\mathrm{combi}}^{\mathrm{SPOKES}} - \beta_{\mathrm{combi}}^{\mathrm{HCA}}$	0.29	2.65	0.3	(-0.48, 1.26)
	$\gamma_{\mathrm{combi}}^{\mathrm{SPOKES}} - \gamma_{\mathrm{combi}}^{\mathrm{HCA}}$	-0.16	1.5	0.55	(-0.66, 0.52)
Criticism	$\alpha_{\mathrm{IY}}^{\mathrm{CPT}} - \alpha_{\mathrm{IY}}^{\mathrm{HCA}}$	0.49	0.34	0.16	(-0.16, 1.18)
	$\beta_{\mathrm{IY}}^{\mathrm{CPT}} - \beta_{\mathrm{IY}}^{\mathrm{HCA}}$	0.06	0.12	0.58	(-0.19, 0.28)
	$\gamma_{\mathrm{IY}}^{\mathrm{CPT}} - \gamma_{\mathrm{IY}}^{\mathrm{HCA}}$	0.19	0.14	0.17	(-0.12, 0.43)
	$\alpha_{\mathrm{combi}}^{\mathrm{SPOKES}} - \alpha_{\mathrm{combi}}^{\mathrm{HCA}}$	0.31	0.4	0.37	(-0.47, 1.17)
	$\beta_{\mathrm{combi}}^{\mathrm{SPOKES}} - \beta_{\mathrm{combi}}^{\mathrm{HCA}}$	-0.09	0.13	0.55	(-0.46, 0.07)
	$\gamma_{\mathrm{combi}}^{\mathrm{SPOKES}} - \gamma_{\mathrm{combi}}^{\mathrm{HCA}}$	-0.12	0.13	0.32	(-0.34, 0.17)

The test results show no significant between-trial difference in any mediation effect ($\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$) for any intervention (IY or COMBI) at the 5% level for either the warmth or the criticism mediation model. The EMO ($\boldsymbol{\beta}$) is expected to be constant across trials, as it represent the mechanism by which the mediator affects the outcome. The treatment effects, ETM ($\boldsymbol{\alpha}$) and DE ($\boldsymbol{\gamma}$), are also constant across trials. This might be because all the factors that potentially cause the treatment effects' trial heterogeneity have been adjusted in the model. The trial heterogeneity test results indicate that given the same intervention, we can set up a more parsimonious model that assumes that mediation effects (and total treatment effects) do not vary with trials. Consequently, the model can be simplified to estimate the mediation effects: $\alpha_{\mathrm{IY}}, \beta_{\mathrm{IY}}, \gamma_{\mathrm{IY}}$; $\alpha_{\mathrm{combi}}, \beta_{\mathrm{combi}}, \gamma_{\mathrm{combi}}$ and $\alpha_{\mathrm{lit}}, \beta_{\mathrm{lit}}, \gamma_{\mathrm{lit}}$. It should be borne in mind that

the literacy only intervention has only been provided in the HCA trial, so that it was not possible to assess any effect heterogeneity in treatment effects for this intervention.

Thirdly, after performing the set of tests described in the previous two steps, we have a simplified IPD meta-mediation model for warmth and criticism mediation analyses respectively. I will then run the simplified models using the IV-MI-BT approach and get the point estimates and the BT distribution of the effects of interest ($\alpha_{\mathrm{IY}}, \beta_{\mathrm{IY}}, \gamma_{\mathrm{IY}}$; $\alpha_{\mathrm{combi}}, \beta_{\mathrm{combi}}, \gamma_{\mathrm{combi}}$ and $\alpha_{\mathrm{lit}}, \beta_{\mathrm{lit}}, \gamma_{\mathrm{lit}}$) for both warmth and criticism. It is also my interest to investigate whether the effects of different interventions are the same. To assess this formally, I test a further set of hypotheses: $\alpha_{\mathrm{IY}} = \alpha_{\mathrm{combi}} = \alpha_{\mathrm{lit}}$, $\beta_{\mathrm{IY}} = \beta_{\mathrm{combi}} = \beta_{\mathrm{lit}}$, and $\gamma_{\mathrm{IY}} = \gamma_{\mathrm{combi}} = \gamma_{\mathrm{lit}}$ using the novel non-parametric bootstrap multi-parameter testing approach proposed in this chapter. The analysis results of the tests are shown in Table 5-4.

Table 5-4 Mediation effects of interest between-intervention variability tests

Mediators	Hypothesis	MD value	95% threshold	p-value
Warmth	$\alpha_{\mathrm{IY}} = \alpha_{\mathrm{combi}} = \alpha_{\mathrm{lit}}$	2.38	6.68	0.29
	$\beta_{\mathrm{IY}} = \beta_{\mathrm{combi}} = \beta_{\mathrm{lit}}$	0.39	6.83	0.80
	$\gamma_{\mathrm{IY}} = \gamma_{\mathrm{combi}} = \gamma_{\mathrm{lit}}$	0.16	6.84	0.92
Criticism	$\alpha_{\mathrm{IY}} = \alpha_{\mathrm{combi}} = \alpha_{\mathrm{lit}}$	0.87	5.95	0.58
	$\beta_{\mathrm{IY}} = \beta_{\mathrm{combi}} = \beta_{\mathrm{lit}}$	1.24	6.49	0.51
	$\gamma_{\mathrm{IY}} = \gamma_{\mathrm{combi}} = \gamma_{\mathrm{lit}}$	0.74	7.52	0.61

The three parenting interventions (IY only, literacy only, IY and literacy combined) are fairly similar. They all include similar components for improving parenting practice with comparable intervention durations and delivered in the same format. As expected, no significant differences were detected over three interventions for the mediation effects of interest for warmth and criticism. Therefore, the parameters of different interventions are combined together in the final IPD meta-mediation model and the final mediation effects to be estimated here are $\alpha, \beta, \gamma, \alpha\beta \; and \; \gamma + \alpha\beta$ for the warmth and criticism mediation analyses.

5.4.5 Interpretation of the results from the simplified IPD meta-mediation model

The systematic IPD meta-mediation model construction approach performed in the previous section provides more efficient estimators of the mediation effects of interest based on the

empirical evidence. At the same time, the between-trial heterogeneity of the effects of interest (including the confounders) is considered in the analysis when necessary. The final IPD meta-mediation analysis model combined with the IV-MI-BT approach is applied to the pooled data of three parenting trials. The effect of the intervention on the mediator (α), the effect of the mediator on the outcome (β), the indirect effect ($\alpha\beta$), the direct effect (γ) and the total effect ($\gamma + \alpha\beta$) are estimated conditional on all the measured confounders (with or without between-trial heterogeneity) using a model that includes the main effects of trials and reflects the features of individual trial designs. The IV-MI-BT approach relax the assumption of no unmeasured confounding of the mediator-outcome relationship and the distributional assumption of the mediators and the outcome in the presence of missing values. The results of the final IPD meta-mediation analyses using the proposed IV-MI-BT combined approach are listed in Table 5-5.

Table 5-5 Results of IPD meta-mediation analysis for expressed warmth and expressed criticism mediators using the IV-MI-BT method

Putative Mediator	Causal mediation parameter	Estimate	SE	P-value	Bias Corrected 95% BT CI
Warmth	α	0.26	0.13	0.05	(0.03, 0.57)*
	β	-0.22	0.22	0.27	(-0.69, 0.2)
	γ	-0.41	0.12	0.01	(-0.63, -0.18)*
	$\alpha\beta$	-0.05	0.06	0.27	(-0.25, 0.03)
	$\gamma + \alpha\beta$	-0.46	0.09	<0.01	(-0.65, -0.27)*
Criticism	α	-0.29	0.12	0.05	(-0.48, -0.08)*
	β	0.16	0.09	0.11	(-0.08, 0.3)
	γ	-0.55	0.10	<0.01	(-0.78, -0.37)*
	$\alpha\beta$	-0.05	0.04	0.26	(-0.11, 0.03)
	$\gamma + \alpha\beta$	-0.6	0.09	<0.01	(-0.76, -0.41)*

The results of the expressed warmth IV-MI-BT IPD meta-mediation analysis show that the parenting interventions of three trials reduced child antisocial behaviour problems by 0.46 standard deviations, of which 0.05 standard deviations are due to increasing parental expressed warmth and 0.41 standard deviations are due to factors other than expressed warmth conditional on all the measured confounders (baseline child behaviour, baseline warmth, parent depression with trial-heterogeneity, child reading ability, lone parent, child gender and parent education). The results of the expressed criticism IV-MI-BT IPD meta-mediation analysis show that the parenting interventions of three trials reduced child antisocial behaviour by 0.6 standard deviations, of which 0.05 standard deviations are due to

reducing parental expressed criticism and 0.55 standard deviations are due to factors other than expressed criticism conditional on a same set of measured confounders.

Statistically significant results based on 95% biased corrected confidence intervals are asterisked in Table 5-5. The total effect of parenting interventions significantly reduced child antisocial behaviour at 5% level in all three trials. The interventions also significantly changed parental expressed warmth and criticism. However, the magnitude of the causal indirect effects via warmth and criticism are very small and are not significant at 5% level. Although the standard errors of the indirect effects are reduced compared to the single trial (SPOKES) IV mediation analysis in Chapter 4, the small magnitude of the indirect effects requires even smaller standard errors to achieve significance. In summary, the IV-MI-BT IPD meta-mediation analysis using data from three parenting trials improved the precision of the estimates of causal mediation effects via increasing the sample size. Statistically significant total treatment effects and significant direct treatment effects of the parenting interventions on the child outcome were found. However, the synthesised estimates of the indirect effect via parental expressed warmth or parental expressed criticism were small and failed to reach significance at the 5% level.

5.4.6 Comparison between the single trial and pooled trials mediation analyses

For a deeper understanding of the synthesized mediation analysis results of the IV-MI-BT IPD meta-mediation analysis, mediation analysis of individual contributing trials using the IV-MI-BT approach were conducted and the analysis results are listed in Table 5-6.

The analysis results of the SPOKES trial have already been displayed and interpreted in Chapter 4. In SPOKES, the effects of both the treatment on the mediator (α) and the mediator on the outcome (β) are of large/moderate size, which leads to a moderate/small indirect effect ($\alpha\beta$) with standardised estimates equal to -0.15 and -0.12 for expressed warmth and expressed criticism mediators respectively. However, the synthesised mediation analysis using three parenting trials obtained a very small indirect effect. As shown in Table 5-5, the standardised estimates of the indirect effect of expressed warmth and expressed criticism mediators are both -0.05.

Table 5-6 Results of mediation analysis for expressed warmth and expressed criticism mediators using the IV-MI-BT method for each trial individually

Putative Mediator	Trial	Treatment	Causal mediation parameter	Estimate	Bias Corrected 95% BT CI
Warmth	SPOKES	COMBI	α	0.46	(-0.12, 1.22)
			β	-0.33	(-0.83, 2.6)
			γ	-0.38	(-0.92, 0.18)
			$\alpha\beta$	-0.15	(-2.62, 0.12)
			$\gamma + \alpha\beta$	-0.54	(-0.73, -0.28)*
	CPT	IY only	α	-0.69	(-1.33, -0.25)*
			β	0.08	(-0.51, 0.35)
			γ	-1.02	(-1.47, -0.65)*
			$\alpha\beta$	-0.05	(-0.32, 0.25)
			$\gamma + \alpha\beta$	-1.07	(-1.39, -0.71)*
	HCA	COMBI	α	-0.09	(-0.48, 0.59)
			β	0.13	(-0.06, 0.94)
			γ	-0.76	(-3, -0.32)*
			$\alpha\beta$	-0.01	(-0.43, 0.03)
			$\gamma + \alpha\beta$	-0.77	(-3.29, -0.29)*
		IY only	α	0.04	(-0.5, 0.44)
			β	-0.01	(-0.41, 0.56)
			γ	-0.35	(-1.97, 0.95)
			$\alpha\beta$	0	(-0.2, 0.11)
			$\gamma + \alpha\beta$	-0.35	(-1.92, 1.01)
		Literacy	α	-0.15	(-0.49, 0.14)
			β	0.19	(-0.35, 0.86)
			γ	-0.84	(-3.01, 0.61)
			$\alpha\beta$	-0.04	(-0.6, 0.04)
			$\gamma + \alpha\beta$	-0.88	(-3.36, 0.67)
Criticism	SPOKES	COMBI	α	-0.39	(-0.68, 0.16)
			β	0.32	(-0.07, 0.61)
			γ	-0.4	(-0.63, -0.12)*
			$\alpha\beta$	-0.12	(-0.3, 0.09)
			$\gamma + \alpha\beta$	-0.52	(-0.79, -0.2)*
	CPT	IY only	α	-0.63	(-1.12, -0.22)*
			β	0.1	(-0.6, 0.31)
			γ	-0.97	(-1.45, -0.57)*
			$\alpha\beta$	-0.06	(-0.25, 0.3)
			$\gamma + \alpha\beta$	-1.03	(-1.33, -0.61)*
	HCA	COMBI	α	0.08	(-0.61, 1.39)
			β	0.22	(0, 0.6)
			γ	-0.7	(-1.71, -0.3)*
			$\alpha\beta$	0.01	(-0.13, 0.31)
			$\gamma + \alpha\beta$	-0.69	(-1.4, -0.26)*
		IY only	α	-0.27	(-1.07, 0.23)
			β	0.19	(-0.15, 0.52)
			γ	-0.59	(-1.07, -0.09)*
			$\alpha\beta$	-0.06	(-0.4, 0.01)
			$\gamma + \alpha\beta$	-0.65	(-1.15, -0.07)*
		Literacy	α	-0.51	(-2.09, -0.09)*
			β	0.24	(-0.08, 0.52)
			γ	-0.67	(-1.13, -0.26)*
			$\alpha\beta$	-0.12	(-0.5, -0.01)*
			$\gamma + \alpha\beta$	-0.79	(-1.25, -0.31)*

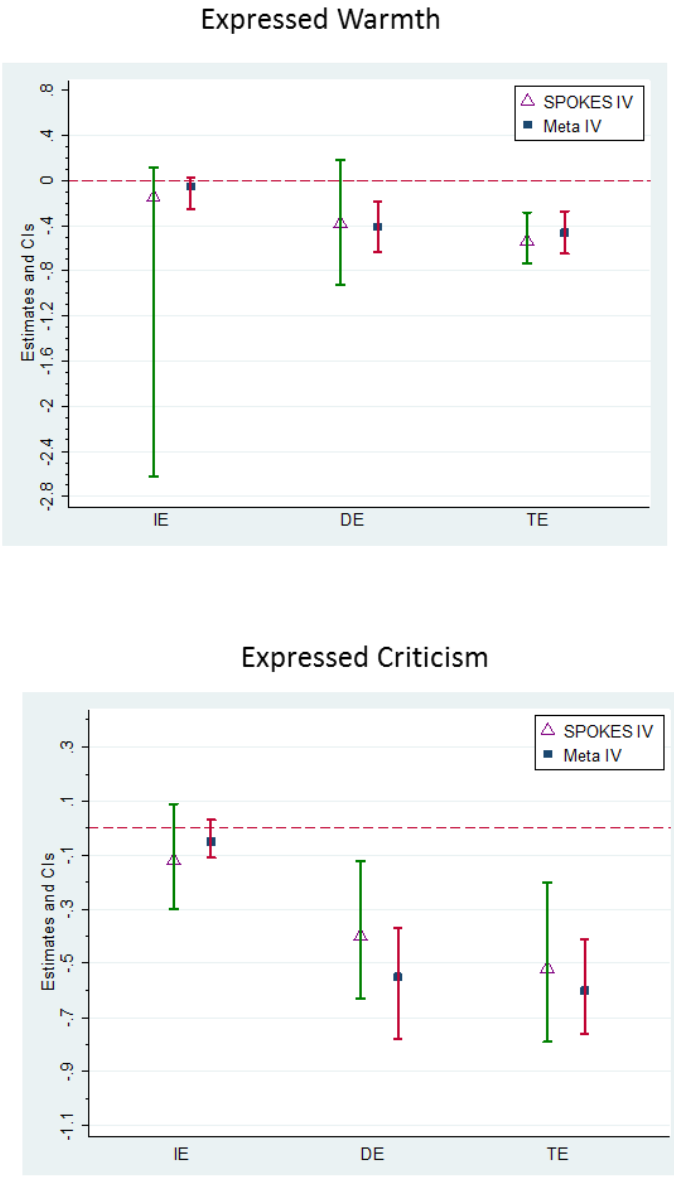
*indicates the significant results based on 95% bias corrected confidence interval

The reduction of the mediation effect can be explained by looking at the mediation analysis results of each trial individually. In CPT, the IY only parenting intervention reduced the expressed criticism but the effect of expressed criticism on child outcome is very small. However, the IY only parenting intervention in CPT did not increase the expressed warmth; on the contrary, it reduced the expressed warmth and there is no evidence of the effect of warmth on the child outcome. In HCA, three active parenting interventions (COMBI, IY only, and Literacy only) were investigated. For expressed warmth, none of the three interventions increased this putative mediator. For expressed criticism, the COMBI intervention did not reduce this putative mediator; the IY only intervention reduced parental expressed criticism and in turn it improved child behaviour but the estimate of the indirect effect was small and not significant at 5% level; interestingly, the results of the literacy only intervention (that also has a component of improving parental behaviour) indicate that this intervention improved child behaviour via reducing parental expressed criticism. Briefly, the three contributing trials include interventions that did not change the putative mediators, so that there is no evidence that a change in the putative mediator leads to a change in the child outcome for these subgroups of the population. Including these population subgroups in the pooled data analysis diluted the estimates of the mediation effects of interest. Therefore, the synthesised mediation effects estimates are small in size compared to the estimates obtained from the single trial (SPOKES) mediation analysis using the IV-MI-BT method.

Of note, the between-trial variations of the treatment effect on the putative mediators were observed, but they are not statistically significant based on the test results obtained in Section 5.4. Therefore, pooling data from the three trials and using the simplified model to estimate mediation effects of interest are appealing and the estimates of mediation effects synthesised are generalised and interpretable. The potential reasons for the observed between-trial variation of the treatment effect on the mediator are: the differences related to the delivery of the interventions (e.g. the people who delivered the interventions, the adherence to the interventions) between trials; the adjustment of the intervention components in different trials (e.g. different numbers of training sessions were designed in different trials; the intervention components may have altered slightly in later trials based on the experiences of previous trials and improved knowledge).

Compared to the SPOKES single trial mediation analysis in Table 4-3, the widths of the confidence intervals of the indirect effect estimates in the pooled data mediation analysis (Table 5-5) are narrower. To illustrate the results, Figure 5-4 plots the estimates and confidence intervals of the causal effects of interest of the SPOKES single trial mediation analysis and the Meta trial IPD mediation analysis using the IV-MI-BT approach. These results provide evidence of the potential benefit of pooling data from multiple trials for IPD meta-mediation analysis using the IV-MI-BT method. The improvement of the precision for the IV mediation approach is due to the increased sample size and the combination of effects between trials based on the results of the effects heterogeneity tests.

Figure 5-4 Results of SPOKES single trial and Meta trial (IPD) mediation analyses for warmth and criticism mediators using IV-MI-BT approach



IE ($\hat{\alpha}\hat{\beta}$): Indirect Effect; DE ($\hat{\gamma}$): Direct Effect; TE($\hat{\gamma} + \hat{\alpha}\hat{\beta}$): Total Effect.

5.5 Discussion

5.5.1 The strengths of the IPD meta-mediation analysis based on the IV-MI-BT approach

Compared with single trial IV-MI-BT mediation analysis, the new IV-MI-BT IPD meta-mediation analysis offers many advantages, including greater power (i.e. larger pooled sample size) and the opportunity to build a more cumulative science (i.e. examining the similarity of the effects across studies and potential reasons for dissimilarities). A summary of the strengths of the IPD meta-mediation analysis approach proposed in this chapter is as follows: Firstly, the IPD meta-mediation analysis uses the information provided by the raw individual-participant data from multiple contributing trials and provides synthesis and efficient estimators of the effects of interest. Therefore, the IPD meta-mediation analysis has advantages in terms of economy, precision and synthesis inference. Secondly, combination of the IV-MI-BT approach and the IPD meta-mediation analysis offers all the advantages of the IV-MI-BT approach to the meta-analysis. The advantages of the IV-MI-BT approach have been discussed in Chapters 3 and 4. Thirdly, the proposed systematic approach to constructing IPD meta-mediation analysis provides an opportunity to investigate further research questions including the between-trial heterogeneity of the effects of interest and the between-intervention variability of the effects of interest. The answers to these research questions lead to a better understanding of the underlying mechanism and potentially inspire the improvement of the interventions. Importantly, the simplified IPD meta-mediation analysis model takes account of the between-trial heterogeneity only when needed, which provides better estimators of the effects of interest and potentially improves the efficiency of the mediation analysis. Finally, the framework of IPD meta-mediation analysis proposed in this chapter can be used in combination with other mediation analyses as appropriate, including the IV mediation analysis using the LIML estimator for weak instruments for reducing the weak IV bias and the MI-BT mediation analysis discussed in Chapter 3 if the assumption of no unmeasured confounding is acceptable.

5.5.2 Important considerations when performing IPD meta-mediation analysis based on the IV-MI-BT approach

We know that the IPD meta-mediation analysis combines the individual participant data from multiple trials. Some trials may have collected the mediation data but not reported it in the primary publication that focused on the intervention's effectiveness, so that the unpublished data may exist in a raw state. Additionally, different trials may measure the

same underlying concept using different instruments. These two example situations make the harmonisation of the measurements across different trials a time-consuming process. Secondly, using different harmonisation methods may affect the analysis results, therefore further investigation needs to be done to assess the impact of the harmonisation on the effects estimate. Thirdly, the limitations of the IVs approach discussed in Chapter 4 are also the limitations of the IV-MI-BT and IPD meta-mediation analysis combined approach. Finally, in the IPD meta-analysis, the trial designs must be appropriately modelled, as failure to do so (i.e. specific randomisation procedure and clustered data) may lead to biased estimates of causal mediation effects. In this context, attention needs to be paid to the interpretation of the causal effects of interest after conditioning on the trial-specific variables necessary to account for specific trial designs (i.e. conditioning on the recruitment cohort in HCA).

Chapter 6 General Discussion

6.1 Summary of methods and findings

Despite the effectiveness of well-established parenting programme interventions on reducing child antisocial behaviour (NICE, 2013), their mediation mechanisms have rarely been formally investigated. The IY parenting programmes are established based on the theory that the intervention increases positive parenting, reduces negative parenting and consequently improves child behaviour (Patterson, 1982, Bowlby, 1969/1982, Scott and Dadds, 2009). This underlying theory reflects the basic idea of mediation: that is, the mediating variable (mediator) transmits the effect of the intervention to the outcome. The goal of causal mediation analysis is to identify mediators and disentangle the total effect of the intervention into direct and indirect effects. This thesis empirically and formally tested the hypothetical mediation mechanism and estimated the causal effects of interest using data from three existing IY parenting programme RCTs, leading to a better understanding of the mediation mechanism of parenting interventions and the potential to improve underlying theory and refine parenting interventions.

In terms of statistical methodology development, firstly this thesis developed the MI-BT approach to handle missing data and relax the distributional assumption of the mediator and the outcome with observed confounders included in the mediation model. To further improve this mediation approach, I developed the IV-MI-BT approach, which relaxed the assumption of no unmeasured confounding between the mediator and outcome relationship in trials via combining the MI-BT approach with the IVs method. Finally, I proposed a systematic approach to conducting IPD meta-mediation analysis that can provide efficient mediation effect estimators in the integrative mediation analysis and answer new research questions using pooled data from multiple existing trials. In the following paragraphs, I will provide a chapter by chapter summary of this thesis.

Chapter 1 reviewed the underlying theory of developing parenting interventions, the effectiveness of parenting interventions as established via RCTs, and the currently performed mediation analyses of parenting interventions. The findings from the literature review brought out the necessity and the importance of conducting formal mediation analysis for parenting interventions. In addition, previous research into parenting

interventions provided the theoretical and empirical foundation for constructing the hypothesised parenting programme mediation models with putative parenting practice mediators (*positive parenting* and *negative parenting*). Chapter 1 also explained mediation-related terminologies and defined causal mediation parameters of interest for preparing the causal mediation analyses presented in the subsequent chapters. At the end of Chapter 1, the statistical challenges of traditional mediation analysis approaches were summarized and identified as the issues to be solved in this thesis. The statistical challenges were as follows:

1. Traditional mediation analyses assume the absence of confounding of the mediator-outcome relationship in order to ensure a causal interpretation. This assumption might be overly simplistic in some RCTs and can lead to biased effect estimates;
2. Missing values might be present in observed confounding variables, putative mediators and clinical outcome variables. A CC mediation analysis is inefficient and might suffer missing data biases;
3. Trials of parenting interventions employ complex trial designs. Failure to take account of design features such as cluster effects is likely to lead to biased estimators of mediation effects;
4. Discrete (not continuous) mediators and/or outcomes make the normality distributional assumption of traditional linear regression mediation implausible. Therefore, statistical inference based on assumption of normality may become subject to bias;
5. Mediation analyses require large sample sizes to achieve sufficient statistical power, and in general the sample size of a single trial is relatively small. Methods for improving the precision of the mediation effect estimates are required.

Chapter 2 introduced three parenting trials, SPOKES, CPT and HCA, in respect of the trial participants, study design, intervention and measurement. These trials all aimed to reduce child antisocial behaviour by improving parenting practices. They targeted similar populations, investigated analogous parenting programmes and employed compatible measures of child and parent outcomes. The sample sizes of SPOKES, CPT and HCA were 112, 110 and 213 respectively. This trial introduction chapter also illustrated that the three trials all have clustered/ hierarchical data structures and missing values are present in the mediators, confounders and outcomes.

Pooling data from multiple trials for an integrative analysis is one of the aims of this thesis. A considerable amount of preparatory work, including data cleaning and measurements harmonization, was carried out in Chapter 2 to make the measurements obtained from multiple trials comparable. Chapter 2 demonstrated the steps for harmonising the measures of baseline demographics, baseline behaviours, parenting practices outcomes and child outcomes. In addition, using the strategies proposed in Chapter 2, I identified a list of representative measures for the positive and negative parenting practices respectively. They are *interview play, expressed warmth*, and *observed positivity* for *positive parenting*; *interview smacking, expressed criticism*, and *observed negativity* for *negative parenting*.

Chapter 3 addressed measured confounding of the mediator-outcome relationship, the presence of missing values, the hierarchical structure of the data and the non-normality of outcomes by proposing a novel approach: the *MI-BT mediation analysis approach*. In this approach, the measured confounders were tested and conditioned in the mediation model to reduce the confounding effect on the mediator-outcome relationship; the MICE technique was applied for handling missing values; multi-level modelling method was used to account for the hierarchical data structure; and the non-parametric bootstrapping approach relaxed the distributional assumption when generating statistical inferences for causal effects of interest. The MI-BT approach provides consistent estimates with a 95% non-parametric bootstrap bias corrected confidence interval for the causal effects of interest, assuming that there is no hidden confounding and under a plausible MAR missing data mechanism. The procedure for combining Multiple Imputation, bootstrapping and a mediation analysis model was demonstrated in Figure 3-2.

This MI-BT mediation analysis approach was applied to the SPOKES trial to test which parenting practice factors mediated (transmitted) the effect of the IY and literacy combined parenting intervention to child antisocial behaviour. The six putative parenting practice mediators proposed in Chapter 2 were tested in single-mediator models separately and seven mediator-outcome confounders (*baseline child behaviour, baseline parenting practice, child's gender, child's reading ability, parent's education, parent's depression* and *lone parent*) were conditioned in all mediation models. The tests identified two mediators of the effect of the parenting intervention on reducing child outcome. They are parental expressed warmth and parental expressed criticism. The results showed that 38% of the total effect is

mediated via reducing parental criticism and 22% of the total effect is mediated via increasing parental warmth respectively. The standardised indirect effect ($\hat{\alpha}\hat{\beta}$) of parental criticism is significant at the 95% level (estimated effect -0.19; 95% bias corrected confidence interval from -0.42 to -0.05). The standardised indirect effect ($\hat{\alpha}\hat{\beta}$) of parental warmth is also significant (estimated effect -0.11; 95% CI from -0.25 to -0.01). The above results provided empirical evidence that the IY and literacy combined parenting intervention reduced child antisocial behaviour via improving parental expressed warmth and reducing parental expressed criticism, assuming that there is no unmeasured confounding between the parenting mediator and the child outcome. These findings are supported by the theory underlying parenting programmes.

Chapter 4 improved the *MI-BT mediation analysis approach* by combining it with the IV methods that relaxed the assumption of no unmeasured confounding between mediator and outcome. The new approach is called the *IV-MI-BT approach*. The application of IV method to mediation analysis was implemented cautiously. The IVs of the endogenous mediator were grounded in theory and, as far as possible, verified based on empirical data. The interactions between treatment randomisation and covariates were used as IVs of the endogenous mediators. In other words, these IVs are actually the moderators of the treatment effect on the mediators. In order to choose the most practically useful IVs, a set of criteria were proposed in Section 4.3.2 for verifying the weak instrument bias and the variance inflation of the IV estimation. It is assumed that the set of IVs have no direct effect on the outcome, which is known as the *exclusion restriction* assumption. The *IV-MI-BT* approach assumes that the IVs satisfy the exclusion restriction and the missing data mechanism is MAR. Similar to Chapter 3, the non-parametric BT method and the MICE technique were also used within this new approach and the programming steps for implementing the proposed *IV-MI-BT* mediation analysis approach were detailed in Section 4.3.3.

In SPOKES, the selected IVs for the expressed criticism mediator are the interaction between intervention and baseline parental depression, the interaction between intervention and baseline parental education, and the therapy groups in the treated arm; the selected IVs for the expressed warmth mediator are the number of sessions attended (%) in the treated arm. The same set of confounders as in Chapter 3 was conditioned in the mediation models for

both criticism and warmth. The results of applying the IV-MI-BT mediation analysis approach to SPOKES data showed that the standardised estimate of the indirect effect ($\hat{\alpha}\hat{\beta}$) is -0.12 with 95% biased corrected confidence interval (-0.3, 0.09) for parental criticism and -0.15 with 95% CI (-2.62, 0.12) for parental warmth. The magnitude of the estimated causal indirect effects remains small (according to Cohen's d criteria for standardised differences, an effect size of 0.2 to 0.3 might be a "small" effect, around 0.5 a "medium" effect and 0.8 to infinity a "large" effect) and they can no longer be shown to be significant at the 5% level based on the confidence interval. The results showed that 27.8% of the total effect is mediated via parental warmth and 23.1% of the total effect is mediated via parental criticism, indicating that the percentages of mediation are not small. The *IV-MI-BT* approach reduced the unmeasured confounding bias of the mediation effect estimates at the cost of losing precision. Compared to the *MI-BT* analysis results presented in Chapter 3, the estimate of the indirect effect via warmth increased a little, and the estimate of the indirect effect via criticism decreased slightly, after correcting the unmeasured confounding bias between the mediators and the outcome. The comparison results imply that the unmeasured confounding bias may not be very large, or the instruments used here are weak so that the bias reduction is limited.

Chapter 5 pooled data from three parenting trials (SPOKES, CPT and HCA) and conducted an *Individual Participant Data* (IPD) meta-mediation analysis in order to provide synthesised estimates of the mediation effects of interest and gain insight into how the effects vary between trials. A systematic approach was developed for constructing an IPD meta-mediation model. It starts by constructing a full IPD model which allows all the model parameters to vary with trial, and then the full model is simplified via a novel bootstrap procedure testing whether restrictions can be imposed to hold some parameters constant across trials. Failure to detect significant between-trials heterogeneity will lead us to assume constant effects across different trials. Under this empirically supported assumption, the full model can be simplified via combining the parameters of different trials to form the final IPD meta-mediation model with the most efficient estimator of the effects of interest. Besides, the results of testing the between-trial heterogeneity in effects provide empirical evidence of the nature of various effects. Chapter 5 also proposed the *IV-MI-BT IPD meta-mediation analysis approach,* which combined the IPD meta-mediation analysis with the *IV-MI-BT* approach in order to provide synthesised estimates of the effects of interest without

assuming the absence of unmeasured confounding and in the presence of missing values in the covariates.

The pooled data set combining the SPOKES, CPT and HCA trials was analysed using the newly developed *IV-MI-BT IPD meta-mediation analysis approach*. The same set of mediator-outcome confounders and the IVs as selected in Chapter 4 were included in the meta-mediation model. The BT and MICE techniques were conducted within each trial separately, allowing the between-trial heterogeneity for all the effects to be tested. For baseline confounders, significant between-trial differences were detected in baseline child reading ability, baseline parental criticism, and baseline parental depression. There is no evidence of between-trial heterogeneity for the effect of the intervention on the mediator (α), the effect of the mediator on the outcome (β) and the direct effect of the intervention on the outcome (γ). The synthesised estimates of the indirect effect of both warmth and criticism are small in magnitude and are not significantly different to zero based on 95% bias corrected confidence intervals. More specifically, the estimates were an indirect effect via parental expressed criticism ($\hat{\alpha}\hat{\beta}$) of -0.05 (95% CI from -0.11 to 0.03; 8% of the total effect) and an indirect effect via parental expressed warmth ($\hat{\alpha}\hat{\beta}$) of -0.05 (95% CI from -0.25 to 0.03; 11% of the total effect).

The results of the mediation analysis for each trial individually are listed in Table 5-6. Looking at the estimates of the direct, indirect and total effects in each trial and each intervention group individually, I found that parenting interventions in some contributing trials did not change the target mediators. As a result, the estimates of the indirect effects are small for these trials and intervention groups. On the other hand, the observed cross-trial variations of the effects of interest were not significant using the test proposed in Chapter 5, so that it is appropriate to combine the effects across trials when conducting an IPD meta-analysis. The inclusion of trials whose intervention did not change the putative mediators diluted indirect effects and led to smaller and insignificant indirect effects via expressed warmth and criticism. Of note, the confidence intervals of the indirect effect estimates of the IPD meta-analysis are narrower than the single SPOKES trial mediation analysis. These results provide evidence for the potential benefit of pooling data from multiple trials for IPD meta-mediation analysis using the IV-MI-BT method. The improvement of the precision for the IV

mediation approach is due to increased sample size and the combination of effects between trials based on the results of the effects heterogeneity tests.

6.2 Strengths and weaknesses of the study

6.2.1 Design and measures

The randomised controlled trial design enables us to study the causal effects of the treatment on the intermediate outcome (mediator) and on the distal outcome. This also benefits the causal mediation analysis, in which the effects of the treatment on the intermediate and distal outcomes are the key components for calculating the causal direct effect, indirect effect and total effect. In other words, randomly assigning participants to the treated and the control groups eliminates confounding of the relationship between the treatment and the mediator or the outcome, provided that the RCT has reasonable sample size. This means that only the potential confounding between the mediator and the outcome needs to be considered and a corresponding assumption to be made in order to obtain a causal interpretation of effect estimates.

The three parenting trials included in this project applied multi-method and multi-informant measures across several domains. The comprehensive measures of putative parenting mediators and child outcome provide a rich data source to investigate the mediation mechanism from various aspects of parenting practices, contributing to find the exact component(s) that are serving as mediator(s).

The similarity of the trial designs and the compatibility of the measures across three parenting trials allow us to pool the data from different trials to conduct an IPD meta-mediation analysis and obtain the estimate of the mediation effects of interest in a synthesised and efficient manner. The RCT design with comprehensive conventional measures of the mediators and the distal outcome not only act as crucial factors allowing rigorous mediation analysis but also contribute to form a high quality database of pooled data that can be used to further research questions, such as how the effect is affected by trial level variables. However, measurement methods and scales often differ across trials and the harmonisation of measures across trials might be time-consuming.

One of the challenges in establishing mediation is that the measurement of the mediator should temporally precede the measurement of the outcome to motivate a causal interpretation of the effect of the putative mediator on the outcome and not *vice versa*. However, in many parenting trials, the mediator and the child outcome are assessed at the same time point. When observing data cross-sectionally (instead of longitudinally), the causal effect of the putative mediator on the outcome must be based on theory or prior research. Inferential assumptions such as the correct specification of causal ordering (temporal precedence) and causal direction are also especially important but often difficult to defend (MacKinnon and Fairchild, 2009). Taking the SPOKES trial mediation analysis as an example, the favourable mediation result for expressed emotion might be an artefact because the possible underlying mechanism is that the improvement of the child behaviour in the unmeasured previous session leads the parent to change their expressed emotion at the measured end point. Future studies need to measure mediators in multiple sessions in order to investigate the longitudinal relationships among treatment, mediator and outcome. Secondly, it takes time for the mediator to exert its effects on the outcome. If mediator and outcome variables are measured at the same time, there may not be enough time for the mediator to affect the outcome. In contrast, if the time lag between the mediator and the outcome measures are too long, the effect size of the mediator on the outcome may fade over time. Further details of the limitations of using cross-sectional data to investigate longitudinal relations have been discussed by Gollob and Reichardt (Gollob and Reichardt, 1991). However, if we assume that the system has reached equilibrium and the direction of any relationship is known from theory, then the cross-sectional snapshot data can reflect the relations accurately.

A major limitation of mediation analysis is that it requires large sample sizes to estimate relevant parameters with adequate precision and detect mediators with sufficient power. Usually, single trials are not powered for mediation analysis, while pooling data from multiple trials to conduct meta-mediation analysis may be the solution. Meta-analysis requires the identification and application of clear inclusion criteria to define target populations, trial designs and appropriate measurement of concepts. Linking the limitation of cross-sectional data analysis with pooled data meta-analysis, it is possible that different contributing trials take different snapshots of the relationships with the observed effect sizes depending on the time lag. In our three parenting programme trials, the post-treatment

measurements of parenting practice mediators and child outcomes were measured at 1 year for SPOKES, at 9-11 months for HCA and at 5-7 months for CPT (see Table 2-1). The observed mediation effect size differences between trials might be due to the time gap between trials. The meta-analysis proposed in this project provided a synthesized estimate of the effects of interest across different time points.

6.2.2 Causal mediation analysis

The traditional mediation analysis approach as proposed by Baron and Kenny (Baron and Kenny, 1986) is based on linear regression and has had strong influence in psychology and social sciences in recent decades. Recently, this traditional approach has been extended to allow for the presence of treatment-mediator interactions in the outcome regression model using counterfactual definitions of direct and indirect effects (VanderWeele and Vansteelandt, 2010). Statistical packages such as *paramed* have been developed to allow non-continuous outcomes and mediators (Emsley and Liu, 2013). The traditional mediation analysis approach and its extension works above assume no confounding between the mediator and the outcome. To relax this restrictive and unrealistic assumption, we provide solutions from two aspects: inclusion of the measured confounders in the models for the mediator and outcome dependent variables; extending IV approaches to the context of parenting intervention trials to address the potential issue of unmeasured confounding. Within this causal mediation inference framework, I still assume linearity of the relationships, reliable and valid measures of the variables, and the absence of mediator-outcome effect modification. Under these assumptions, the causal effects to be estimated are the effect of treatment on mediator (ETM), the effect of mediator on outcome (EMO), the indirect effect (IE), the direct effect (DE) and the total effect (TE).

In this thesis, the IVs we used are the treatment-mediator effect modifiers, or, say, the interactions between treatment randomisation and covariates. They are needed to relax the no unmeasured confounding assumption made by the traditional mediation analysis. For the expressed criticism mediator, the IVs selected are 1) the interactions between treatment and baseline parental depression, 2) the interaction between treatment and baseline parental education, and 3) the therapy groups in the treated arm are the IVs. For the expressed warmth mediator, the number of sessions attended in the treated arm is the IV. Of note, the treatment-mediator effect modifiers are actually of two different types: the

baseline variables (such as parental depression and parental education), and the post-randomisation variables (e.g. therapy groups in the treated arm, and number of sessions attended in the treated arm). The use of interactions between randomisation and baseline variables as instruments has been recommended by researchers (Albert, 2008, Small, 2012, Emsley et al., 2010, Ten Have et al., 2007) to investigate causal mediation. Randomisation ensures that there is no unmeasured confounding for the interaction instrument and the outcome, which offers a promising start as instruments. I assume that the baseline variables influence the size of the effect of the treatment on the putative mediator, but they do not influence the size of the direct effect of the treatment on the outcome. It is important to recognise that the IVs assumptions are made for the interaction terms, but not for the main effect of the baseline variables. Post-randomisation variables like the therapy groups and the number of sessions attended in the treated arm (not available in the control arm) are practically equivalent to the interaction term between treatment randomisation and these post-randomisation variables. However, to fulfil IV assumptions, the main effects (causal or otherwise) of these counterfactual variables (e.g. of one's ability to attend therapy sessions when offered) on the outcome have to be assumed to be zero, which is somewhat restrictive. Therefore, extra caution should be paid when using post-randomisation treatment effect modifiers as IVs for the endogenous mediators. A new way forward to enable valid causal mediation analyses is to generate interaction IVs by design and a recently published article by Imai and colleagues (Imai et al., 2013) has discussed how to identify causal mechanism (mediation) using specific experimental designs.

Although I am assuming that IV assumptions hold for selected IVs on theoretical grounds, estimates constructed from IV analyses can be biased for finite samples. The bias is towards the direction of the confounded association between the endogenous variable (the mediator) and the outcome, and the size of the bias relates to the statistical strength of the association between the instrument and the endogenous variable (Burgess and Thompson, 2011). In this project, the largest F-statistics for the IVs of a single endogenous variable in the first-stage regression is approximately 3, which leads to a relative bias of about 30% (Sawa, 1969, Stock and Yogo, 2002, Cragg and Donald, 1993). Briefly, this means that the bias of the IV estimator is approximately 30% that of the bias of the OLS estimator in the traditional mediation analysis and in the same direction as the OLS bias. However, the extent of the bias of the newly developed ML type IV estimator (2SML) is still not clear and further research is

needed on this topic. It is also known that when the IVs are weak, the IV estimator has a long-tailed distribution (Imbens and Rosenbaum, 2005), which is not well approximated by a normal distribution. One of the strengths of the *IV-MI-BT* method is that we used the non-parametric bootstrap approach to provide statistical inference for the IV estimator without making distributional assumptions. In theory, this method reduces the bias of the statistical inference obtained using normal distribution based methods such as the 2SLS method.

As briefly mentioned in the previous section, the drawback of the IV method is that it requires a large sample size to gain sufficient power in the mediation analysis and provide a precise causal effects estimate. The IPD meta-mediation analysis in this project pooled individual participant data from three parenting trials, using an IV approach to provide synthesised and more precise estimates of the causal mediation parameters of interest. Recently, IPD meta-analysis has been applied for Mendelian randomization (Burgess et al., 2012) using genetic variants as IVs to estimate causal associations from observational data. Although the research topic was not about causal mediation, combining IPD from multiple trials for meta-analysis has the same purpose: that is, to improve the precision of the IV analysis via increasing the sample size. In this project, the putative mediators, confounders and IVs were selected based on the observed trial data. One may argue that empirical predictor variable selection can lead to over-fitting. However, the same variables selected from one trial (SPOKES) were used in the *IV-MI-BT* IPD meta-mediation analysis, which avoided the potential bias due to selecting variables based on the observed data and led to more reliable conclusions. In addition, the framework developed for the IPD meta-analysis took account of the different trial designs and provided efficient causal effect estimates via systematically testing the between-trial heterogeneity of the effects of interest (including the confounding effects). These methods were illustrated using data from parenting trials, but they can be applied to a broader range of causal IV or Baron and Kenny type meta-mediation analysis using individual participant data.

The statistical analyses developed in this project have several novel aspects. This thesis conducted one of the first causal mediation analyses to use a combination of instrumental variable, Multiple Imputation and the bootstrapping method. It was also the first to use the IV method to estimate causal mediation analysis in trials of parenting programmes and conducted the first meta-mediation analysis using IPD. In detail, this thesis developed the IV

method in the context of RCTs of parenting programmes for the causal interpretation of the mediation effects without having to assume the absence of unmeasured confounding of the mediator-outcome relationship. Compared to complete case analysis, MI can greatly improve the efficiency of the analysis. As shown in the SPOKES trial, MI led to an increase of more than 50% in the sample size compared with CC. Handling missing values using MI in the setting of causal mediation analyses is a major practical contribution of this project. The non-parametric bootstrap approach provides statistical inferences without making distributional assumptions about the mediators and the outcome. Combination of all three methods for mediation analysis theoretically provides less biased estimation of the causal mediation effects of interest. Additionally, the newly developed *IV-MI-BT* IPD meta-mediation analysis framework using IPD from multiple parenting trials constructed the synthesised and potentially more precise (efficient) estimation of the effects.

6.2.3 Alternative Approaches to Causal Mediation Analyses allowing for hidden confounding

In addition to the IV approach, alternative causal mediation approaches have been developed to estimate mediation effects without assuming that there is no hidden confounding. One of the alternative causal modelling approaches for continuous outcomes is a semi-parametric approach which is known as a structural nested mean model (SNMM) using G-estimation in the context of estimating the causal effects of treatment received in RCTs (Goetghebeur and Lapp, 1997, Goetghebeur and Vansteelandt, 2005). Ten Have et al. (Ten Have et al., 2007) have used a similar G-estimation approach based on *rank preserving models* (RPM), combining the work of Robins and Greenland (Robins and Greenland, 1994) on direct effects of randomized intervention effects for survival outcomes and the work of (Ten Have et al., 2004) on intervention non-adherence. Compared with the SNMM, an additional assumption of RPM is that there is a baseline covariate that interacts with random assignment in predicting the mediator but that does not modify the direct causal effects of the random assignment and the mediator on the outcome. This RPM assumption is equivalent to the assumption that the interaction between baseline covariate and random assignment is an instrumental variable (IV). Extended from the traditional randomisation tests, G-estimation involves a mapping of the observed outcome of each of the participants allocated to the treatment group to the potential treatment-free outcome by subtracting the estimated linear combination of parameters and observed values of intervention, mediator

and baseline covariates for the purpose of making the treatment-free outcome independent of randomisation. The G-estimation procedure produced asymptotically unbiased estimators of the direct and the indirect effects and corresponding standard errors without making the no hidden confounding assumption, relying on these interaction assumptions. Dunn and Bentall (Dunn and Bentall, 2007) show that 2SLS estimation of an SEM with the baseline covariate interaction with random assignment acting as an IV procedure is essentially equivalent to G-estimation of the RPM.

The propensity scores (PS) method offers a potential alternative estimation technique for mediation analysis with different assumptions from those of traditional mediation analysis (Jo et al., 2011). The assumption of no hidden confounding between mediator and outcome made by the traditional mediation analysis is necessary for causal interpretation of the association estimate (β) between mediator and outcome. This assumption also means that across treatment groups, individuals with the same mediator value have the same characteristics and thus the treated and control individuals with the same mediator value can be compared. In fact, as observed in trials, the treated and control individuals who have the same mediator value actually have different characteristics. In this method, the propensity score is used to compare individuals in the treatment and control groups who would have had the same value of the mediator had they been assigned to the same treatment condition. The mediation effects are estimated within strata defined by potential mediator values under treatment and control conditions. However, this method requires a binary mediator and dichotomising a continuous mediator to a binary mediator leads to loss of information. In addition, the PS approach assumes that the principal stratum membership is independent of the potential outcomes given the observed covariates, which is untestable. The PS approach also assumes that the observed covariates are sufficient to identify the stratum membership, which might be arguable, especially when only a few covariates were observed.

6.3 Contributions to the field of parenting intervention

Parenting programme mediation analysis plays a key role for understanding the mechanism by which the parenting intervention improves the child outcome. By identifying the mediator(s) of the parenting intervention, mediation analysis can contribute to improving

the theory underlying parenting intervention, refine complex parenting programmes and potentially enhance the effect of the parenting intervention on child behaviour.

The findings of the *MI-BT* mediation analysis of SPOKES trials with 112 randomised participants indicate that both positive parenting (parental expressed warmth) and negative parenting (parental expressed criticism) mediated the parenting programme treatment effect on child antisocial behaviour. The parental expressed warmth mediator is also supported by the research findings of (Gardner et al., 2006) on RCTs with 120 low-income two- to three-year-old toddlers. It was suggested that the change of the positive parenting (a combined code including constructive activities for the child and the use of positive discipline strategies) mediated the effect of the parenting intervention on improving child behaviour. However, negative parenting (a combined code included criticizing, threatening, using sarcasm, and yelling) did not appear to be a mediator in Gardner's trial, while its mediation effect was significant in the SPOKES trial. It has been suspected that negative exchanges may be more normative at age two and only come to reflect more entrenched (Tremblay et al., 1999), damaging cycles of coercive interaction when the child is older. The appearance of a negative parenting mediation effect in older children (SPOKES children are five to six year old) supports this theory. The appearance of the negative parenting effect might also be due to our analysis adjusting (conditioning in the SPOKES analysis model) for the observed confounding between negative parenting and child behaviour. The conditional no hidden confounding assumption is weaker (less restrictive) than the no confounding assumption made in many substantive applications and we advocate conditioning on observed confounders in mediation analysis. A recent publication (Kling et al., 2010) of the traditional mediation analysis results of a Swedish parent management training (PMT) RCT with parents of 159 children aged three to ten suggested that negative parenting (summary scores of harsh and inconsistent parenting) and positive parenting (summary scores of praise and incentives) are the mediators of the effect of PMT on child antisocial behaviour. The SPOKES trial, the toddler trial and the PMT trial mediation analysis findings indicate that: 1) parenting interventions independently improved positive parenting and reduced negative parenting, and consequently reduced child antisocial behaviour, and 2) child behaviours can be improved by changing different subsets of positive and/or negative parenting practices.

This is the first time that an IV method in combination with MI and bootstrapping, the so-called *IV-MI-BT* method, has been developed and used in parenting programme mediation analysis, which relaxed the assumption of no unmeasured confounding between the parenting practice mediator and the child antisocial behaviour outcome. Compared to the corresponding mediation analysis results using the improved traditional method (*MI-BT* method), the IV estimation of the indirect effect via increasing expressed warmth is larger and the effect via reducing expressed criticism is smaller. These results suggest that mediation analysis that ignores unmeasured confounding is likely to lead to an overestimation of the indirect effect of expressed criticism and an underestimation of the indirect effect of expressed warmth; however, this conclusion is not assured because the direction and the quantity of the bias are also affected by the effect of the included confounders and the validity of the IVs. The IVs selected for warmth and criticism mediators are weak, with F-statistics equal to 3.1 and 1.2 respectively, indicating that the IV estimates reduced the bias of traditional mediation analyses for warmth and criticism by 70% and 17% respectively. As a result of the trade-off between bias reduction and precision, the confidence intervals of the estimates in the SPOKES *IV-MI-BT* mediation analysis are wider than in the *MI-BT* mediation analysis. To further increase the precision and reduce the bias of causal mediation estimation for parenting programmes in future research, the two key elements are: 1) trials with much larger sample sizes, and 2) IVs that are valid and predictive.

The findings of the IPD meta-mediation analysis involving three parenting trials support neither parental expressed warmth nor criticism as the mediator of the parenting intervention on reducing child antisocial behaviour. The results of the IPD meta-mediation analysis are synthesized results of three trials, each of which plays an important role in the meta-mediation analysis considering their relatively even sample size. The mediation of parenting intervention via warmth and criticism may be found in one trial, but this may not occur in the others, which leads to the disappearance of the mediation in a broader setting. Considering the findings of both the SPOKES single trial mediation analysis and the IPD meta-mediation analysis, several discussion points are listed as follows. Firstly, it might be suspected that the positive findings in one trial might happen purely by chance. However, the large/medium size and proportion of the mediation suggests that this is less likely to be due to chance. Secondly, parental expressed warmth and criticism can mediate the effect of the parenting intervention on child outcome in moderate/large proportion, but they are not

the universal and compulsory elements to be changed by the parenting intervention in order to achieve its effect on improving child behaviour. A possible way forward is to model multiple mediators using data from multiple trials. Thirdly, parenting programmes were designed to improve multiple aspects of parenting practices and the effect of the parenting intervention might be mediated via the combination of multiple parenting practice components. To test this theoretical hypothesis, multiple mediators should be modelled simultaneously. The challenges of the multi-mediator model are that it requires knowledge about the causal relationship between mediators and finding valid multiple IVs for multiple mediators might be difficult.

6.4 Future research

Several points in relation to possible future research have already been mentioned in the previous sections. In this section, I will provide a summary of suggestions for future research relating to parenting programmes and methodology for mediation analysis.

Well-designed parenting intervention RCTs with large sample sizes provide opportunities to understand the mechanism of parenting intervention on child antisocial behaviour, test hypothesised mediators, identify the most effective components of the intervention, and in turn improve interventions and refine underlying theory. The SPOKES mediation analysis suggests that improvements in parental expressed warmth and reductions in criticism mediate the effect of the parenting intervention on child antisocial behaviour. As discussed in the previous section, the mediation may be caused by multiple mediators and so each represents a combined effect of multiple correlated parenting practices. Future research needs to identify the independent components of parenting practices that can form hypothesised latent mediators (e.g. positive parenting and negative parenting), and test the mediation contribution of this set of putative mediators using data sets of larger sample size (potentially using IPD from more trials).

Moreover, based on what we have learnt from this project, several recommendations are made relating to the design and measurement of parenting programme studies to investigate causal mediation. Firstly, randomisation is important for studying causal relationships. Although treatment randomisation cannot resolve the potential confounding issue between mediator and outcome, it leads to the causal interpretation of the effect of

the treatment on the mediator and the total effect of the treatment on the outcome. Therefore, treatment randomisation is advocated for causal mediation analysis of parenting intervention. Secondly, trials of parenting interventions are mostly designed to test the effectiveness of the intervention but not to investigate the mediation. In order to achieve sufficient power for testing potential mediators, sample size should be calculated at the design stage accounting for mediation analysis. Thirdly, all the possible confounding variables should be measured prior to randomisation so that the observed confounders can be conditioned in the mediation analysis in order to minimize the confounding bias. Next, temporal precedence is a crucial point in the identification of the mediating process. For the case of parenting programme mediation, it means that the parenting mediator measure should be assessed prior to the child outcome measure, in order to avoid reverse causation. Finally, in terms of measurements, standard measurement methods are recommended to measure mediators and outcomes, as this can make the potential data pooling from multiple trials for meta-analysis more straightforward. If an underlying concept relates to multiple standard measurement methods with different formats (e.g. questionnaire, interview, and direct observation) and different informants (e.g. parent reported, child reported, and teacher reported), the measures using multiple standard methods can contribute to taking measurement error into account.

The instrumental variables selected in the parenting programme mediation analysis are weak instruments that can lead to biased and imprecise effect estimates. Details of the disadvantages of weak instruments have been discussed in Chapter 5. Finding strong and valid IVs is still a big challenge for future research into IV causal mediation analysis for parenting programme RCTs. In this thesis, the treatment effect modifiers are considered as IVs and the effect modifiers are tested on the basis of both theory and empirical work. Further down the path, moderation analysis of the effect of the treatment on the mediator should be considered at the design stage (Landau, 2011) and carried out in more single trials and in pooled trials in order to identify effect modifiers that might serve as IVs for causal mediation analysis. In the case of multi-mediator mediation models, multiple IVs are needed to assure causal interpretation. However, finding multiple valid IVs for multiple endogenous variables might be challenging.

The limitations of cross-sectional analysis have been discussed in Section 6.2.1 of this chapter. As a supplement to cross-sectional analysis, longitudinal data contain more detailed information that provides opportunities to investigate the mediation mechanism over time and across participants. This longitudinal causal mediation area is drawing more and more research interest. In order to maximally exploit all the information collected in the trials and model all mediator and outcome data of longitudinal repeated measures provided by modern trial design, further developments are required in modelling for factor structures and longitudinal data.

Appendix I Data User Guide of Trials of Parenting Programmes

SPOKES Data User Guide

Published Paper

- **Title:** Randomised controlled trial of parent groups for child antisocial behaviour targeting multiple risk factor: the SPOKES project
- **Authors:** Stephen Scott, Kathy Sylva, Moira Doolan, Jenny Price, Brian Jacobs, Carolyn Crook, and Sabine Landau
- **Resource:** Journal of Child Psychology and Psychiatry 51:1 (2010), pp48-57
- **Intension:** To investigate the effectiveness and feasibility of the population bases intervention targeting ineffective parenting, conduct problems, ADHD symptoms and low reading ability that are the risk factors predicting poor outcome.

Trial Protocol

Design

Stage one: screening of all children in the school year.
Stage two: Randomized Controlled Trial (RCT) of eligible cases.

Study population

The trial was named Supporting Parents On Kids Education in Schools (SPOKES) and ran from 1999 to 2001 in 8 schools in Lambeth, London, among the 5% most deprived English Boroughs. All children in reception and year one classes (kindergarten) were screened (n=936).

Eligibility

1. Children had to exhibit conduct symptoms above the screen cutoff level.
2. Parents had to show
    a. ability to understand English;
    b. ability to attend at group times;
    c. interest in attending;
    d. acceptance of RCT study;
    e. child free of clinically apparent developmental delay.

Consent: Written consent was obtained; the local research ethics committee approved the project.

Randomization

Two annual cohorts were screened in four schools, one in the remainder (total 12 cohorts in 8 schools). After screening, 8-16 cases (mean 10.7) per cohort were assessed and then the trial coordinator forwarded cases to the trial statistician who, blind to any other information, randomized them individually to the intervention or control group using GENSTAT. Parents of 112 high scores were randomised to intervention (n=61) or control group (n=51).

Masking

Assessors and parents were blind to allocation status at initial assessment. At follow-up, questionnaires were entered by data staff blind, videotapes were coded by researchers blind, and interviews were carried out by assessors blind.

Parenting group intervention and control

Intervention

Groups: 4-8 parents
Frequency and duration: 2.5 hours one morning per week for 28 weeks
Intervention programmes: 12-week child behavioural programme - based on “Incredible Years” (IY; Webster-Stratton) school age programme; followed by 10-week literacy programme - based on Pause Prompt Praise (PPP; McNaughton, Glynn, & Robinson); finally 6-week revision
Control: Telephone helpline

Measurements

Screen

Teachers and parents were asked to complete the conduct problems scale of the Strengths and Difficulties Questionnaire (Goodman, 2001), with five questions scored not true=0, somewhat true=1, certainly true=2 (range0-10). Additionally the eight DSM IV (American Psychiatric Association, 1994) oppositional-defiant disorder items were used, scored not true=1, just a little true=2, pretty much true=3, very much true=4 (range

0-32). Parent and teacher scores were summed. The cutoff was SDQ >=5 or DSM >=10, one standard deviation above the population mean for 5-6 year olds, designed to capture most cases at risk of lifetime-persistent antisocial behaviour.

Participant characteristics

An interview covered family structure and income, housing type, delivery and developmental history, ethnicity and parental education; the General Health Questionnaire 12 covered maternal psychiatric symptoms (Goldberg et al., 1997).

Measurement Time Points

Time Point 1: Before randomization
Time Point 2: Short telephone interview
Time Point 3: One year after randomization (four months after the end of the experimental intervention)

Parenting

Observation: The procedure of the Conduct Problems Prevention Research Group (CPRG) (Conduct Problems Prevention Research Group, 1999) was used, with videotaping of parent-child interaction for 15 minutes across three tasks: (i) child directed play, (ii) parent directed task, (iii) parent instructs the child to tidy away the toys. Scoring was frequency counts by three raters blind to allocation status; coders used a modified version of the CPRG scheme. Factor analysis gave three summary codes: a. total attends to child; ICC (Intraclass Correlation Coefficient) on 20 tapes was 0.82. b. seek cooperation (question requests in conditional tense, eg "would you tidy the toys away?"), ICC 0.69; c. total commands, ICC 0.83.

Interview: we used a semi-structured interview developed by Michael Rutter and colleagues (Conduct Problems Prevention Research Group, 1999, Bierman et al., 1999, Woodward et al., 1997). Reliability between the three interviewers was calculated on 30 interviews after two months of training on pilot study cases; intraclass correlations ranged from 0.62 to 0.77.

Expressed emotion (EE): this is a measure of emotions expressed towards the child throughout the interview. It was rated on a 5 point scale using Camberwell Family Interview criteria (Vaughn, 1989); for warmth the ICC was 0.76, for criticism 0.73.

Questionnaire: the Parenting Practices questionnaire (Webster-Stratton et al., 2008) has four subscales: positive involvement, appropriate discipline, inconsistent parenting, and harsh discipline; the first two and last two were combined.

Child antisocial behaviour

The Parent Account of Child Symptoms (Taylor et al., 1986) was the trial's primary outcome. This is a standard investigator-based interview similar to, but shorter than the Child and Adolescent Psychiatric Assessment, and has been used in large surveys. Antisocial behaviours (lying, stealing, tantrums, rudeness, disobedience, destructiveness, aggressiveness) are scored 0-3 for severity and frequency in the last month and the mean calculated (range 0-6); ICC was 0.89.

Oppositional defiant disorder diagnosis was elicited from the parent interview using DSM IV criteria (ICC 0.85).

The Eyberg Child Behaviour Inventory (Boggs et al., 1990) is a parent-completed questionnaire of 36 oppositional behaviours. Teachers rated antisocial behaviour using DSM IV questionnaire items.

Child ADHD symptoms

These were measured with the PACS interview; ICC was 0.81.

Child reading ability

This was measured using the British Ability Scale (Elliott et al., 1996a). This is an individually administered test of the child's ability to read single words. Researchers received extensive training until they reached 95% agreement. Assessors were blind to allocation status.

Child emotional disorder symptoms

These were measured by the PACS interview and covered depression, fears, eating and sleeping problems (ICC 0.78).

Participant satisfaction Questionnaire (Webster-Stratton et al., 2008)

Reading Measurements reported in the BJEP paper (Sylva et al., 2008)

Child literacy outcomes

- British Picture Vocabulary Scale (BPVS) (Dunn et al., 1997), which is designed to establish a child's level of receptive vocabulary and to provide some indication of general ability.
- Rhyme and Alliteration, two subscales of the phonological awareness assessment (Bryant and Bradley, 1985). Rhyme and alliteration were combined and are referred to as phonological awareness hereafter.
- Concepts about Print and Writing/Dictation (this last only at post-test, as many children were not able to write at pre-test). The assessments concepts about print and writing/dictation are part of Marie Clay's battery (Clay, 1993).

Experienced researchers received extensive training until they reached a satisfactory level of reliability prior to the actual testing of the children. In addition, they were blind to the parents' group allocation.

Parent reading with child

Parent Account of Child Symptoms (PACS) Semi-structured Interview obtained information on the amount of time mother reads with her child in a week and the use of reading strategies. To create one overall variable of parents' use of different reading strategies, information on the three variables (Scene setting, Emotional encouragement and Literacy strategies) was summed to a total of different strategies used.

Reading information collected but not reported in the papers

- My Child's Leading Questionnaire were used to assess mother rated confidence in schooling; involvement with school; ambition about child literacy; belief in child literacy competence.
- Weekly Reading Diary collected the time read together and the name of books; the time played reading game together and name of games

Note 1: Parent daily report of child behaviours (42 questions for 1 week, the variable name indexed time-point, day and questions number) and Parenting Stress Index (my experience of being a parent – 36 questions and my health and wellbeing – 12 questions) were also employed. The information was not reported in paper but the data were available.

Note 2: Prorating is applied on the questionnaire sub-score calculation. Given the direct observation parenting outcome scores were fully collected for each subject if the assessment was performed, prorating is not used to observation parenting scores.

Note 3: The direct observation parenting outcome derived dataset is saved as "SPOKES Obs data analysis_18Apr2012". Child Global Functioning, Child Enjoyment with Play and Child Social Responsiveness were collected in the raw data but not included in the derived database as they are not considered as putative mediators.

Table 1: SPOKES Project Measurement Scale Summary

Measurement Name	Measurement Subject	Completer/Performer	Scale	Range	Validity
Study Screen					
Strengths and Difficulties Questionnaire (SDQ)	Screening Eligibility	Parent & Teacher	0=true 1=somewhat true 2=certainly	0-10	
DSM IV Questionnaire	Screening Eligibility	Parent & Teacher	1=not true 2=just a little true 3=pretty much true 4=very much true	0-32	
Participant Characteristics					
SPOKES Demographic Information Interview	Family structure and income, housing type, delivery and developmental history, ethnicity and parental education	Parent			
General Health Questionnaire 12 (GHQ)	Maternal psychiatric symptoms	Parent			
Parenting Assessment					
Conduct Problems Prevention Research Group (CPPRG)	Videotape Observation Parent-child interaction: Positive attention; Seek cooperation; Give commands.	Rater	Frequency counts		Total attention - ICC 0.82 Seek cooperation - ICC 0.69 total commands - ICC 0.83
Rutter and colleagues semi-structured interview	Parenting Interview: Play; Praise, Rewards; Consequences; Time Out; Harsh discipline.	Interviewer	Five rating points for six scales		3 interviewers - ICC 0.62 to 0.77
Camberwell Family Interview	Emotions expressed towards child interview: Warmth; Criticism	Interviewer	Five rating points		Warmth - ICC 0.76 Criticism - ICC 0.73
Parenting Practices Questionnaire	Parenting practices questionnaire: Appropriate and positive parenting; Harsh and inconsistent parenting.	Parent			

Child Outcome Assessment					
Parent Account of Child Symptoms (PACS) Semi-structured Interview	Child antisocial behaviours: lying, stealing, tantrums, rudeness, disobedience, refusal bed, destructiveness, aggressiveness; Child ADHD symptoms; Child emotional disorder symptoms	Parent	0-3 for severity and frequency	0-3	anti-social behaviour - ICC 0.89 ADHD - ICC 0.81 Emotional scale - ICC 0.78
DSM IV Interview	Child ODD	Parent			ICC 0.85
Eyberg Child Behaviour Inventory Questionnaire	36 oppositional behaviours	Parent			
DSM IV Questionnaire	Child antisocial behaviours	Teacher			
Reading Assessment					
British Ability Scale (BAS)	Individual administered test on child reading ability	Assessor			95% agreement
Reading Assessment not used in the JCPP paper					
British Picture Vocabulary Scale (BPVS)	Raw score, standardised score, % rank, age equivalent in month	Assessor			satisfactory level of reliability
Phonological awareness test	Rhyme, Alliteration	Assessor			satisfactory level of reliability
Marie Clay's battery of test	Concepts About Print, Writing/dictation	Assessor			satisfactory level of reliability
Parent Account of Child Symptoms (PACS) Semi-structured Interview	Reading time	Parent	minutes per day x no of days/week		
Parent Account of Child Symptoms (PACS) Semi-structured Interview	Reading strategies: Scene setting Emotional encouragement Literacy strategies	Parent	0-2 for strategy level	0-6	
My Child's Learning Questionnaire	Confidence in schooling; Involvement with school; Ambition about child literacy; Belief in child literacy competence	Parent	1-7 for level from low to high	1-7	
Weekly Reading Diary	Read together Play reading games together Total all reading activities	Parent	Mean minutes per day		

CPT (VTST) Data User Guide

Published Paper

- **Title:** Multicentre controlled trial of parenting groups for childhood antisocial behaviour in clinical practice
- **Authors:** Stephen Scott, Quentin Spender, Moira Doolan, Brian Jacobs, Helen Aspland
- **Resource:** British Medical Journal, 323(7306), 28 JULY 2001, 194-197
- **Intension:** To investigate whether a behaviourally based parenting programme would be effective in everyday NHS practice, with standard referrals to child mental health services and regular clinic staff to carry out the intervention.

Trial Protocol

Design: Controlled trial allocation by date of referral.

Study Population: The trial took place from 1995-1999 in four NHS child and adolescent mental health services: Croydon, Brixton/Belgrave/Camberwell, St George's (all South London); Chichester (West Sussex).

Eligibility criteria: Children aged 3 - 8 years who were referred for antisocial behaviour to their local multidisciplinary child and adolescent mental health service (n=430).

Exclusion criteria:

Children were clinically apparent major developmental delay, hyperkinetic syndrome, or any other condition requiring separate treatment.

Parents had to be able to understand English and attend at group times.

Consent: Written consent was obtained. The relevant ethics committees approved the project. All the eligible children with consent and assessed (n=141)

Assignment

- The allocation was determined by date of receipt of referral letter (avoid the bias from referrers, clinic staff and parents).
- Overall ratio of intervention vs. Control: 2.06:1 (35 intervention blocks: 17 control blocks)
- Block size: Minimum 6 cases.
- The sequence of the block is determined non-randomly for each centre annually in advance

Masking

- Parent were blind to allocation at the initial assessment (informed location after first assessment)
- The interviews were carried out by researchers blind to the duration or sequence of blocks.

Parenting group intervention and control

Intervention

Parenting Programme: Webster-Stratton basic videotape parent training programme

Group size: 6 - 8 parents

Frequency and duration: 2 hours each week for 13-16 weeks

Control: No intervention (waiting list)

Measurements

Participant characteristics

Parent Account of Child Symptoms (PACS) covered family structure and income, housing type, delivery and developmental history, ethnicity and parental education.

Measurement time points

1. Parents on entry to the trial (Before)
2. After completion of the intervention/waiting list period (After)
3. Five to seven months later (Follow up)

Efficacy measurements: Six measures of child behaviour (1 interview, 4 questionnaires and ICD-10 for conduct disorder diagnosis); One of parent behaviour (Direct observe)

Parenting

Direct observation of parenting: An 18 minute structured play task was given to mother and child at home and videotaped. 20 cases were randomly selected and coded using a manual21 by a rater blind to their status. Parental praise and inappropriate commands were counted and combined to give a ratio; intra-class correlation coefficients were 0.96 and 0.97 respectively.

Semi structured interview of parenting: Michael Rutter and colleagues (Conduct Problems Prevention Research Group, 1999, Bierman et al., 1999) developed a semi-structured interview that the frequency of the withdrawal of child's privileges, the child is praised or rewarded, "timeout" punishment implementation or

harsh discipline are measured in single item and scored on a scale of 0-4. A mean score of twelve items measuring the frequency of the parent participating in play activities with the child over the course of the week (including weekend) on a scale of 0 to 2 and further 2 subscales relating to the frequency of time the parent spends playing with the child through the week measuring creative and non-creative play.

Expressed parenting emotion: this is a measure of emotions expressed towards the child throughout the interview. It was rated on a 5 point scale using Camberwell Family Interview criteria (Vaughn, 1989).

Child antisocial behaviour

Interview:

Parent Account of Child Symptoms (PACS) interview was the primary outcome measure. It is a well-validated semi-structured interview that uses investigator–based criteria to assess the frequency and severity of antisocial behaviours such as fighting, destruction and disobedience, and emotional symptoms. The Kappa inter-rater reliability statistic on 20 randomly selected interviews was 0.84 for the conduct problems scale, 0.81 for the hyperactivity scale and 0.76 for the emotional problems scale.

Questionnaire:

1. Strengths and Difficulties Questionnaire
2. Child Behaviour Checklist
3. Parent Defined Problems: the parent lists the three problems they would most like to see change, and indicates the severity of each on a ten-centimetre line labelled 'not a problem' at one end and 'couldn't be worse' at the other.
4. Parent Daily Report: 36 behaviours are recorded as present or absent each day for a week. This measure is widely used as an alternative to prolonged direct observation in the home by an independent observer.

A diagnosis of conduct disorder (oppositional-defiant type) was made if ICD 10 research criteria were met at interview.

Table 2: CPT (VTST) Project Measurement Scale Summary

Measurement Name	Measurement Subject	Completer/Performer	Scale	Range	Validity
Participant Characteristics					
PACS Demographic Information Interview	Family structure and income, housing type, delivery and developmental history, ethnicity and parental education	Parent			
Parenting Assessment					
Parenting behaviour observation	Videotape Observation Parent-child interaction: Praise Inappropriate command	Rater	Ratio of frequency counts over combined		Praise - ICC 0.96 Inappropriate commands - ICC 0.97
Parenting behaviour interview	Frequency of the withdrawal of child's privileges, the child is praised or rewarded, "timeout" punishment implementation or harsh discipline, creative play, non-creative play	Rater	Mean scale		
Parenting expressed emotion	emotions expressed towards the child throughout the interview	Rater	0-5 interval scale		
Measurement Name	Measurement Subject	Completer/Performer	Scale	Range	Validity
Child Outcome Assessment					
Parent Account of Child Symptoms (PACS) Semi-structured interview	Child antisocial behaviours: such as fighting, destruction and disobedience Child ADHD symptoms Child emotional disorder symptoms	Parent	0-3 for severity and frequency	0-3	Antisocial behaviour - ICC 0.84 ADHD - ICC 0.81 Emotional scale - ICC 0.76
Strengths and Difficulties Questionnaire (SDQ)	Child antisocial behaviours	Parent			
Child Behaviour Checklist (Questionnaire)	Child antisocial behaviours	Parent			
Parent Defined Problems (Questionnaire)	Child antisocial behaviours	Parent			
Parent Daily Report (Questionnaire)	36 Child antisocial behaviours	Parent			

HCA Data User Guide

Publication

Title: Which type of parenting programme best improves child behaviour and reading? The Helping Children Achieve trial

Authors: Stephen Scott , Kathy Sylva ,Celia Beckett, Moira Doolan, Angeliki Kallitsoglou, Jeni Beecham & Tamsin Ford, with the HCA study teams

Resource: UK Department for Education, Research Team Final Report

Intension: To assessing the effectiveness of three different parenting programmes to reduce anti-social behaviour and improve reading, in primary school children living in an inner disadvantaged London Borough and a South West city.

Trial Protocol

Design: Randomized Controlled Trial (RCT) with multiple arms.

Stage 1 (screening): Children in reception, year one, or year two were screened or referred for anti-social behaviour by parents and teachers who completed a brief questionnaire, described below. Children whose scores on this questionnaire showed elevated levels of anti-social behaviour were then assessed for eligibility by an interview with the parent.

Stage 2 (randomisation): Families who met the eligibility criteria and said they were interested in taking part in the study were assessed on a range of detailed measures. Eventually 213 families took part, they were randomly assigned to the four intervention and control groups: “Incredible Years” (IY) (n=56); A programme to improve literacy (LIT) (n=53); IY-Literacy (COMBI) (n=50); Signposting (Control) (n=54).

Stage 3 (12 weeks measurement): Brief measures to monitor progress and see which aspects of parenting and child behaviour were beginning to change (mediators) were assessed 12 weeks after the start of the intervention.

Stage 4 (End of study measurement): Recruited families were assessed in detail again, 9-11 months after the first assessment, to assess whether there is a sustained improvement in outcomes.

Study population

The trial was named Help Child Achieve (HCA) and ran February 2008 and March 2012. 2655 families with children aged 5-7 in an inner disadvantaged London Borough (Hackney) and a South West city (Plymouth) were screened or referred to assess levels of child disruptive behaviour.

Eligibility

1) Children met the screen cut-off: based on either the conduct problems (antisocial behaviour) scale on the Strengths and Difficulties Questionnaire (SDQ) or the Diagnostic Statistics Manual (DSM) oppositional defiant scale.
2) Parent’s ability to speak functional English.
3) Interest in taking part in the study.
4) Child score equal or above 0.7 on the Parent Account of Child Symptoms, Disruptive Behaviour scale.
5) Child free of global developmental delay.
6) Child score equal or above 70 on the British Picture Vocabulary Scale, a test related to general cognitive ability(Dunn et al., 1997).

Consent: Written consent was obtained; the local research ethics committee approved the project.

Randomisation

Participants were recruited in blocks and the recruited participants were randomised to one of four intervention groups by an independent statistician. The recruitment block is related to the recruiting year and month. The general rules were that batches should be determined within a month of the case being eligible for the trial. The availability of the intervention groups varies for different blocks. There was a four-year overall plan laying out which interventions would be available and the plan was determined prior to any cases being randomised. The ratio of the participants between intervention groups may differ from one-to-one in order to ensure that the total numbers were approximately equal across groups. The ratio is associated with the recruiting year and month, and the location (Hackney or Plymouth).

Masking

Assessors and parents were blind to allocation status at initial assessment. At follow-up, questionnaires were entered by data staff blind, videotapes were coded by researchers blind, and interviews were carried out by assessors blind.

Parenting group intervention and control

Interventions

The interventions offered were:

1) A literacy-based intervention programme that helps parents support their child's reading (LIT).

The programme combines the Pause Prompt Praise(McNaughton et al., 1987) approach to reading with a 'whole language' approach focusing on meaning(Sylva et al., 2011). This group train programme lasts for 10 weeks, 2-hour sessions, including a home visit and a family literacy workshop, and an additional 2 sessions on how to help their child to concentrate and not be oppositional during shared reading.

2) A well-established parent-child relationship programme that targets behaviour (IY);

The "Incredible Years" Parent Group programme (Webster-Stratton and Hancock, 1998) lasts 12 weeks and each session is 2 hours. The first 6 weeks concentrate on how to build positive relationships and promote desirable child behaviour and constructive activity through play, praise and rewards. The play element focuses on sensitive response to the child and parental approval of child on-task behaviour. The second 6 weeks focus on handling misbehaviour, including ignoring minor misbehaviour, establishing positive routines, applying consequences, and using 'time-out'.

3) A combination of both these two programmes (COMBI);

Families allocated to the combined programme were offered the "Incredible Years" programme followed by the SPOKES literacy programme; the total number of sessions offered was 22.

4) A signposting service that provides parents with information about where to get help (Control).

The control group participated in a Signposting and Information service. Parents were provided with a telephone helpline, which identified appropriate services for parents' concerns about their child and informed them about how to access these services.

Measurements

Screen

Teachers and parents were asked to complete the conduct problems scale of the Strengths and Difficulties Questionnaire (Goodman, 2001). Conduct problems: disobedience, lying, fighting, stealing and temper were scored on a scale of 0-2 (not true=0, somewhat true=1, certainly true=2) and ranged 0-10.

Additionally the eight DSM IV (American Psychiatric Association, 1994) oppositional-defiant disorder items including anger, loosing temper, arguing, deliberately annoying others, refusing to comply, spiteful and vindictive behaviour, blaming others and being argumentative were used, and scored on 0-2 scale from no problems to a frequent problem.

The cutoff was SDQ >=3 or DSM >=5, one standard deviation above the population mean for 5-6 year olds, designed to capture most cases at risk of lifetime-persistent antisocial behaviour.

Educational special needs. The parents were asked in the screen whether the child had any special educational needs and what help they were getting for them. This was categorised as a dichotomous variable of those in receipt of extra help or not. This measure was used as a covariate of outcomes.

Parent and Teacher Reported Child Reading Ability. As part of the screen parents and teachers were asked to report on reading ability on a scale of 1 (cannot read yet) to 6 (reads very well). These questions were repeated at the mediator stage 12 weeks after the intervention.

Socio-demographic

Measures of the families' socio–demographic characteristics were collected using a semi-structured interview used in SPOKES (Scott et al., 2010b) trial which included details of the family structure, occupation (used to assess the socio economic status) and whether the child receives free school meals.

Socio Economic Status: Details of parents' employment was assessed using the National Statistics Socio-Economic Classification (analytic class)(Office for National Statistics, 2005).The resulting data was categorised into four groups as there was an uneven distribution amongst the sample with a higher proportion of SES VIII. The four final groups were I- II: managerial or professional; III-V: intermediate, small employers, supervisory; V-VII: lower routine, technical and routine posts; VIII: never worked or unemployed.

Parental education: This data was collected at interview and covered the mother's educational qualifications, categorised into three groups where 1 = "educated to 16yrs, 2 = "educated to 18+/secretarial/technical qualification and 3 = "educated to degree level or professional or teacher training or degree not finished".

Ethnicity: Parents were also asked for details of their ethnicity based on the ONS categories(Office for National Statistics, 2002). The original 16 point distribution was reduced to a 2 point scale of White British or ethnic minority due to small number of individual ethnic groups.

Parental Mental Health

Depression: The moods and feelings questionnaire DASS (Depression, Anxiety, Stress Lovibond & Lovibond, 1995) comprises items measuring depression, anxiety and stress. Each of these scales are made up of 7 items scored on a scale of 0 to 3 where 0 = "Not true", 1 = "True to some degree", 2 = "Considerably True" and 3 = "Very true", giving a total possible score of 21 for each scale, and where a higher score represents a greater degree of depression, anxiety and/or stress experienced by the respondent.

Aggression: Index of Marital Satisfaction (You and Your Partner) – Verbal and Physical Aggression. (Corcoran & Fischer, 2000) The Questionnaire is a 16 item scale measuring the relationship of the mother of the child (or father if main care giver) and her partner. Each item is measured across a scale of 1 = "Never", 2 = "Sometimes" and 3 = "Often" and asks questions relating to both positive and negative interactions within the relationship. The scale then produces two subscales divided into 7 items measuring verbal aggression (scores ranging from 7 to 21) and 3 items measuring physical aggression (scores ranging from 3 to 9).

A total index of marital satisfaction score is compiled by reverse scoring and summing the remaining items that are not included in the verbal and physical aggression subscales. The result gives a value between 6 and 18 with the scale giving a negative score regarding the degree of satisfaction within the relationship (i.e. a higher score = poorer relationship/more marital dissatisfaction).

Measurement Time Points

Time Point 0: Screen
Time Point 1: Pre-assessment (before randomization)
Time Point M1: 6 weeks post intervention
Time Point M2: 12 weeks post intervention
Time Point 2: Post assessment (9-11 months after the start of the intervention)

The two pre and post assessment points (Time Point 1 and Time point 2) are measured similarly for three parenting trials (CPT(VTST), SPOKES and HCA) and the measures from these two assessment points are used for meta-mediation analysis across three trials. Time Point M1 and M2 are only measured in a subset of the HCA sample, so that they are not included in the mediation analysis.

Measure of Parenting

Interview: we used a semi-structured interview developed by Michael Rutter and colleagues (Conduct Problems Prevention Research Group, 1999, Bierman et al., 1999, Woodward et al., 1997). The frequency of the withdrawal of child's privileges, the child is praised or rewarded, "timeout" punishment implementation or harsh discipline is measured in single item and scored on a scale of 0-4. A mean score of twelve items measuring the frequency of the parent participating in play activities with the child over the course of the week (including weekend) on a scale of 0 to 2 and further 2 subscales relating to the frequency of time the parent spends playing with the child through the week measuring creative and non-creative play.

Expressed emotion (EE): this is a measure of emotions expressed towards the child throughout the interview. It was rated on a 5 point scale using Camberwell Family Interview criteria (Vaughn, 1989); for warmth the ICC was 0.76, for criticism 0.73.

Reading time and strategies interview (Sylva et al., 2008) this measure provides an indication of the time the parent spends with the child reading and the strategies that they use to create the right environment and to help the child with difficulties. The overall time was worked out from the number of times a week the parent spent with the child reading multiplied by the minutes spent. The different strategies for enabling a positive atmosphere and appropriate support for reading were summed from the five questions each scored 0-2.

Observation: The procedure of the Conduct Problems Prevention Research Group (CPRG) (Conduct Problems Prevention Research Group, 1999) was used, with videotaping of parent-child interaction for 15 minutes across three tasks: (i) child directed play, (ii) parent directed task, (iii) parent instructs the child to tidy away the toys. Scoring was frequency counts by three raters blind to allocation status; coders used a modified version of the CPRG scheme.

Questionnaire: The Alabama Parenting Questionnaire (APQ) (Shelton et al., 1996) is a 15 item scale measuring parenting behaviour, consisting of five subscales made up of 3 items each; "Positive Parenting", "Inconsistent Discipline", "Poor Supervision", "Involvement" and "Corporal Punishment". Each item is measured across a scale of 1 to 5 where 1 = "Never", 2 = "Almost never", 3 = "Sometimes", 4 = "Often" and 5 = "Always", giving a total possible score ranging between 3 and 15 for each of the five subscales. As well as the five conventional APQ subscales, two additional subscales have been created summing the two positive subscales ("positive parenting" and "involvement") and the three negative subscales ("inconsistent discipline", "poor supervision" and "corporal punishment") to create a total positive subscale and total negative subscale respectively. Higher scores on all five conventional subscales as well as the overall total positive and negative scales represent a greater degree of that particular factor regardless of whether the scale is slanted positively or negatively.

Parents' view of the study: Twelve weeks after the intervention the parents were asked for their views of the trial, their confidence in managing the child's behaviour now and in the future and any changes they saw in the child's behaviour and reading ability. Parents are asked to show on six-point scale (1=very unconfident to 6=very confident) how confident they felt in managing their child's behaviour.

Child Antisocial Behaviour

The Parent Account of Child Symptoms (Taylor et al., 1986) is the primary outcome of the study. It is a standard investigator-based semi-structured interview. The measure was used to assess the severity and frequency of the child's disruptive behaviour through assessing detailed accounts of several common situations. The questions include lying, stealing, tantrums, rudeness, disobedience, destructiveness, aggressiveness, features of antisocial behaviour in children of this age. The 8 items are each rated for severity (0-3) and frequency (0-3) on a 4-point scale. The mean score of all 8 items is computed to yield the total disruptive behaviour score (ICC 0.89).

Visual Analogue Scale (Aitken, 1969) provides the opportunity for parents to report the nature and intensity of their child's difficulties that is concerning them most on a 10 cm scale and for this to be compared at later time points for the same problem. It was administered by questionnaire at the pre-assessment, 12 weeks after the intervention and at the post assessment.

The Eyberg Child Behaviour Inventory (Boggs et al., 1990) consists of 36 items designed to assess parent-reported conduct problems, and measures the frequency with which problems occur (Intensity Score) as well as the number of problems. This questionnaire has very well established validity. This measure was collected at the pre and post assessment stage of the trial.

Child ADHD symptoms

Child symptom of attention deficit hyperactivity disorder (ADHD) was assessed via PACS that measures the severity and frequency of the child's restless and inattention.

Child literacy outcomes

Child reading ability: This was measured using the British Ability Scale (Elliott et al., 1996a). This is an individually administered test of the child's ability to read single words. Researchers received extensive training until they reached 95% agreement. Assessors were blind to allocation status.

British Picture Vocabulary Scale (BPVS) (Dunn et al., 1997), which is designed to establish a child's level of receptive vocabulary and to provide some indication of general ability. Concepts about Print and Writing/Dictation (this last only at post-test, as many children were not able to write at pre-test): The assessments concepts about print and writing/dictation are part of Marie Clay's battery (Clay, 1993). Wechsler Individual Achievement Test (WIAT) (Wechsler, 2005) assesses the academic achievement of children, adolescents, college students and adults, aged 4 through 85. The test enables the assessment of a broad range of academics skills or only a particular area of need. The WIAT-II is a revision of the original WIAT (The Psychological Corporation), and additional measures. There are four basic scales: Reading, Math, Writing, and Oral Language. Within these scales there are a total of 9 sub-test scores. Elementary reading attitude survey (PAWS) (Kear et al., 2000) was used to measure the recreational reading score and academic reading score. Phonological Assessment Battery (PHAB) (Frederickson et al., 1997). This is a practical measure for identifying pupils with significant phonological difficulties. PHAB comprises six standardised tests including alliteration, naming speed, rhyme, spoonerisms, fluency, and non-word reading tests.

Table 3: HCA Project Measurement Scale Summary

Measurement Name	Measurement Subject	Completer/Performer	Scale	Range	Validity
Study Screen					
Strengths and Difficulties Questionnaire (SDQ)	Screening Eligibility	Parent & Teacher	0=true 1=somewhat true 2=certainly	0-10	
DSM IV Questionnaire	Screening Eligibility	Parent & Teacher	0=true 1=somewhat true 2=certainly	0-16	
Participant Characteristics					
HCA Demographic Information Interview	Family structure and income, housing type, social economic status, delivery and developmental history, ethnicity and parental education	Parent			
The moods and feelings questionnaire DASS (Depression, Anxiety, Stress)	Parental depression, anxiety and stress	Parent			
Index of Marital Satisfaction (You and Your Partner) – Verbal and Physical Aggression.	Parental aggression	Parent			
Parenting Practice Assessment					
Conduct Problems Prevention Research Group (CPPRG)	Videotape Observation Parent-child interaction: Positive attention; Seek cooperation; Give commands.	Rater	Frequency counts		Total attention - ICC 0.82 Seek cooperation - ICC 0.69 total commands - ICC 0.83
Rutter and colleagues semi-structured interview	Parenting Interview: Play; Praise, Rewards; Consequences; Time Out; Harsh discipline.	Interviewer	Five rating points for six scales		3 interviewers - ICC 0.62 to 0.77
Camberwell Family Interview	Emotions expressed towards child interview: Warmth; Criticism	Interviewer	Five rating points		Warmth - ICC 0.76 Criticism - ICC 0.73
The Alabama Parenting Questionnaire (APQ)	Positive Parenting, Inconsistent Discipline, Poor Supervision, Involvement and Corporal Punishment	Parent	Five rating points	3-15	

Parent Reading Time and Strategies					
Parent Account of Child Symptoms (PACS) Semi-structured Interview	Reading time	Parent	minutes per day x no of days/week		
Parent Account of Child Symptoms (PACS) Semi-structured Interview	Reading strategies: Scene setting Emotional encouragement Literacy strategies	Parent	0-2 for strategy level		
Child Antisocial Behaviour Assessment					
Parent Account of Child Symptoms (PACS) Semi-structured Interview	Child antisocial behaviours: lying, stealing, tantrums, rudeness, disobedience, refusal bed, destructiveness, aggressiveness; Child ADHD symptoms; Child emotional disorder symptoms	Parent	0-3 for severity and frequency	0-6	anti-social behaviour - ICC 0.89 ADHD - ICC 0.81 Emotional scale - ICC 0.78
Visual Analogue Scale	nature and intensity of their child's difficulties	Parent	10 cm scale		
Eyberg Child Behaviour Inventory Questionnaire	36 oppositional behaviours	Parent			
Child Literacy Outcome Assessment					
British Ability Scale (BAS)	Individual administered test on child reading ability	Assessor			95% agreement
British Picture Vocabulary Scale (BPVS)	Raw score, standardised score, % rank, age equivalent in month	Assessor			
Marie Clay's battery of test	Concepts About Print, Writing/dictation	Assessor			
Wechsler Individual Achievement Test (WIAT)	Reading, Math, Writing, and Oral Language	Assessor			
Elementary reading attitude survey (PAWS)	recreational reading, academic reading, recreational reading and academic reading	Assessor			
Phonological Assessment Battery (PHAB)	alliteration, naming speed, rhyme, spoonerisms, fluency, and non-word reading	Assessor			

Note:
Mother's mental health variables are standardised using general population mean and standard deviation for the purpose of combining three trials (CPT(VTST), SPOKES and HCA) together. Mother's mental health at baseline is one of the putative baseline moderators for our parenting trials moderation study. It was measured in the three parenting trials using different instruments.

In CPT(VTST), BECK Depression Inventory 21(BDI-21) (Beck At, 1961) was employed. BDI-21 includes 21 questions (items).The scoring method of each item is in ordinal 0-1-2-3, therefore the total (sum) score ranged 0-63.

In SPOKES, General Health Questionnaire -12 (Goldberg, 1972b) GHQ-12 was used. GHQ-12 consists of 12 items and each item has four degrees/levels - (less so than usual, no more than usual, rather more than usual and much more than usual). There are two main scoring methods of GHQ-12.

(1) Likert score
(2) GHQ score (after the name of the questionnaire)

An Example:

Have you recently been feeling sad and gloomy?	Colum1	Colum2	Colum3	Colum4
	Less so than usual	No more than usual	Rather more than usual	Much more than usual
Likert Score	0	1	2	3
GHQ Score	0	0	1	1

The total (sum) score range is 0-36 using Likert scoring method and 0-12 using GHQ scoring method. In SPOKES the Likert score is applied.

The HCA trial used Depression Anxiety Stress Scales 21(DASS-21) measuring mother's mental health in three dimensions (depression, anxiety and stress). The DASS-21 consists of three 7-itme self-report scales that measure depression, anxiety and stress. A 4-point severity scale (0-1-2-3) measures the extent to which each state has been experienced over the past week. The total DASS-21 score ranged 0-21 for each dimension.

The standardised score is calculated using the formula:
(raw score – population mean) / Population standard deviation

For CPT(VTST) Beck-21 standardised score, the general population mean and standard deviation is 7.25 and 5.85 respectively that are taken from a published literature (van Hemert et al., 2002).
For SPOKES GHQ-12 standardised score, the general population mean and standard deviation is 11.50 and 5.08 respectively that are taken from a published literature (Pevalin, 2000).
For HCA DASS-21 standardised score, the general population mean and standard deviation is 2.83 and 3.87 respectively that are taken from a published literature (Henry and Crawford, 2005).

Data correction:

The data is investigated by data manger and it is confirmed that there are 213 participants in HCA trial. Participant ID-2355 was both allocated to literacy and attended on the 2009 May literacy and then dropped out after one session from both the intervention and the study. Participant ID-2700 was randomised to signposting.

Exploration of Direct observation data in the CPT(VTST), SPOKES and HCA trials

Stephen Scott, ninth of January 2 013

I examined the observational data for each trial and in a merged data set with reasonable number of the HCA cases

Factor analysis of average scores across all three tasks

This gave three factors. The first, with 29% of the variance was Alpha commands, don't, impossible, and negative effect; the second with 19% of the variance was facilitate, positive attending, and beta commands; the third was seeking cooperation.

Thus along with alpha commands go quickly given commands and prohibitions; all of these are parents trying to exert control, which is perhaps why it runs with negative effect. Other than alpha, there not a very good way of giving control, according to theories. Facilitate and positive attending go together fairly obviously. Seeking cooperation one might have expected to load negatively with the quick commands or positively with the second factor but did not. On theoretical grounds, it would seem likely that positive parenting should be represented by positive attends, negative parenting by negative effect, and limit setting by some proportion between collaborative control and less effective commands (for example, negative commands, i.e. beta commands and impossible commands).

Intervention effects on parenting across all three trials

<u>positive parenting</u>

total attends average p = .015 attends free play p =.007

facilitation no effects

total attends average confirmed as mediator test

<u>Limit setting</u>

Seek cooperation average p = .003, free play no effect

Alpha commands average no effect, tidy up p= .15

Total directives no effect

Negative commands no effect

on theoretical grounds, seek cooperation divided by total negative commands chosen as mediator test

<u>negative parenting</u>

Negative affect average P equals .09, free play equals .017

No real justification for choosing free play alone for this variable

parenting negative effect confirmed as mediator test

<u>Subsequent elementary mediation tests</u>

In my simple Baron and Kenny model of mediation, using residual of mediator change in a regression either with residual outcome change as dependent, and mediator change and treatment as independent, or outcome time 3 as dependent and outcome time one and mediator change and treatment as dependent, both ways get no mediation effect. This is despite some effects of the mediator on the outcome, i.e. path b.

Trial and task differences

most of this analysis was to look at commands, where we are uncertain about best mediator, but as there were few differences on commands between trials; differences on positive parenting and negative parenting were not be explored. There were considerably higher rates of commands and negativity in the CPT(VTST). The Lego task got considerably higher rates of commanding and facilitation

Appendix II STATA ®12 programming code

MI-BT approach mediation analysis using SPOKES trial

Part 1: Generating point-estimate of the parameters of interest

****Give the values of computer directory path, number of imputation and seed to global variables path, number of imputations and seed respectively*

```
gl path = "U:\PhD Parenting Trial Analysis\SPOKES"
gl nimpute=20
gl seed=542168375
gl alpha=0.95
local confvar = "mumdep1 childiq1 loneparn gender pareducn"
local medvar = "eecriticism"
```

*****Step 1: Apply Multiple Imputation by chained equation to the original data**

****ice command - Multiple Imputation by Chained Equations*

****The variables in the first row of ice command are child behaviour outcome, directly observed parenting mediators at time point 2*

****The variables in the rows below ice are*

**1. child behaviour at baseline, directly observed parenting behaviours at baseline and*

**2. baseline characteristics to be included in the analysis model*

**3. the dummy grouping variables*

**4. the alternative measurement of child behaviour at baseline and time point 2 and the auxiliary baseline characteristics variables*

**5. the interview parenting behaviour variables at baseline and time point 2*

**6. the alternative interview parenting behaviour variables at baseline and time point 2*

**7. the questionnaire parenting behaviour variables at baseline and time point 2*

**8. the expressed emotion parenting behaviour variables at baseline and time point 2*

****Use cmd option to define the regression command to be used in imputation - linear regression is used here*

****Use match option to do predictive mean matching for discrete variables*

****Use eq option to define customised prediction equations for each incomplete variable*

****Predictor variables for child outcome of interest (i.e pacscon2):*

**1. baseline characteristics (i.e. age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn)*

**2. baseline child outcome and alternative child outcome at baseline and time point2 (i.e. pacscon1 pacshyp1 pacshyp2)*

**3. group variables (i.e. i.interven `bgrp`ba'' `tgrp`th'')*

**4. parenting behaviour variables at time point 2 measured via interview, questionnaire, direct observation and EE*

** (i.e. ivplay2 vsmack2 ivconseq2 quposit2 quneg2 qulimit2 obposit2 obneg2 oblimit2 eewarmth2 eecriticism2)*

```
***Predictor variables for parenting behaviour at baseline:
        *1. baseline characteristics (i.e. age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn for ivplay1)
        *2. child outcome at baseline and time point2 (i.e. pacscon1 pacscon2 for ivplay1)
        *3. batch variables (i.e. i.rawbatch for ivplay1. As treatment randomization is applied to parenting trials
        *   and the baseline measurements were done before receiving treatment, we assume the conditional distribution of
        *   baseline variables do not depend on intervention group variable and therapy group variables given the observed values)
        *4. the other parenting behaviour variables measured by the same measurement method (i.e. ivsmack1 ivconseq1 for ivplay1) at baseline and
        *   all the parenting behaviour variables measured by the same measurement method (i.e. ivplay2 ivsmack2 ivconseq2 for ivplay1) at time point 2
        *5. variables measured the same parenting behaviour but by different measurement methods at baseline
        *   (alternative parenting behaviour at baseline. i.e. obposit1 quposit1 eewarmth1 for ivplay1)

***Predictor variables for parenting behaviour mediator (time point 2)
        *1. baseline characteristics (i.e. age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn for ivplay2)
        *2. child outcome at baseline and and time point2 (i.e. pacscon1 pacscon2 for ivplay2)
        *3. group variables(i.e. i.interven `bgrp`ba'' `tgrp`th'' for ivplay2)
        *4. all the parenting behaviour variables measured by the same measurement method (i.e. ivplay1 ivsmack1 ivconseq1 for ivplay2) at baseline and
        *   the other parenting behaviour variables measured by the same measurement method (i.e. ivsmack2 ivconseq2 for ivplay2) at time point 2
        *5. variables measured the same parenting behaviour but by different measurement methods at time point 2
        *   (alternative parenting behaviour mediators. i.e. obposit2 quposit2 eewarmth2 for ivplay2)
***The prediction equations for parenting behaviour variables below follow the same structures as ivplay1 and ivplay2
***The variables below are baseline characteristics variables (i.e. childiq1 mumdep1 pareducn freemeal loneparn ethminor)
***The predictor variables are:
  *1. the other baseline characteristics variables
  *2. child outcome at baseline and time point2 (i.e. pacscon1 pacscon2)
  *3. batch variables (i.e. i. rawbatch. As treatment randomization is applied to parenting trials
  *   and the baseline measurements were done before receiving treatment, we assume the conditional distribution of
  *   baseline variables is not depend on intervention group variable and therapy group variables given the observed values)
  *4. parenting behaviour variables at baseline (i.e. ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1)
***The same imputation model as the one of the bootstrap samples is applied here
***The method is valid under MAR. The variables that are allowed to drive missingness under MAR are those included in the imputation step.
  use "$path\SPOKES_Short_03Apr2013", clear

   ***Set the therapy group value to 0 all the control arm observations and calculate dummy therapy group variables for later Multiple Imputation
     replace therapgp=0 if interven==0
     tabulate therapgp, gen(thergroup)
     local theralist thergroup3 thergroup4 thergroup5 thergroup6 thergroup7 thergroup8 thergroup9 ///
                     thergroup10 thergroup11 thergroup12 thergroup13 thergroup14
```

```
***Set the batch values to another variable for generating batch dummy variables for both treated and control arms
  gen rawbatch=randombatch

***Set batch variable to zero to all the intervention observations and generate dummy batch variables for the control arm for later Multiple Imputation
  replace randombatch=0 if interven==1
    tabulate randombatch, gen(batchgroup)
    local bacthlist batchgroup3 batchgroup4 batchgroup5 batchgroup6 batchgroup7 batchgroup8 batchgroup9 ///
                    batchgroup10 batchgroup11 batchgroup12

  ice pacscon2 obposit2 obneg2 oblimit2 ///
    pacscon1 obposit1 obneg1 oblimit1 age i.gender childiq1 loneparn freemeal mumdep1 ///
    i.interven i.rawbatch `bacthlist' `theralist'  ///
    pacshyp1 pacshyp2 ethminor pareducn ///
    ivplay1 ivsmack1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
    ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
    quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
    eewarmth1 eecriticism1 eewarmth2 eecriticism2, ///
    cmd(ivplay1 ivplay2 ivsmack1 ivsmack2 ivconseq1 ivconseq2 ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
        eewarmth1 eewarmth2 eecriticism1 eecriticism2 childiq1 mumdep1 pacshyp2 ///
        quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 obposit1 obneg1 ///
        oblimit1 obposit2 obneg2 oblimit2:regress) ///
    match(ivplay1 ivplay2 ivsmack1 ivsmack2 ivconseq1 ivconseq2 ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
         eewarmth1 eewarmth2 eecriticism1 eecriticism2) ///
    eq(pacscon2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
                 pacscon1 pacshyp1 pacshyp2 ///
                 i.interven `bacthlist' `theralist' ///
                 ivplay1 ivplay2 ivsmack2 ivconseq2 ///
                 quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
                 obposit1 obneg1 oblimit1 obposit2 obneg2 oblimit2 ///
                 eewarmth1 eecriticism1 eewarmth2 eecriticism2, ///
       ivplay1:  age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
                 pacscon1 pacscon2 ///
                 i.rawbatch ///
                 ivsmack1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
                 obposit1 quposit1 eewarmth1, ///
       ivplay2:  age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
                 pacscon1 pacscon2 ///
                 i.interven `bacthlist' `theralist' ///
                 ivplay1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
                 obposit2 quposit2 eewarmth2, ///
       ivsmack1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
```

```
          pacscon1 pacscon2 ///
          ivplay1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
          obneg1 quneg1 eecriticism1, ///
ivsmack2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          ivplay1 ivsmack1 ivconseq1 ivplay2 ivconseq2 ///
          obneg2 quneg2 eecriticism2, ///
ivconseq1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          ivplay1 ivsmack1 ivplay2 ivsmack2 ivconseq2 ///
          oblimit1 qulimit1, ///
ivconseq2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          ivplay1 ivsmack1 ivconseq1 ivplay2 ivsmack2 ///
          oblimit2 qulimit2 eecriticism2, ///
ivpraise1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          ivsmack1 ivconseq1 ivpraise2 ivsmack2 ivconseq2 ///
          obposit1 quposit1 eewarmth1, ///
ivpraise2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          ivpraise1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
          obposit2 quposit2 eewarmth2, ///
ivrewards1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          ivsmack1 ivconseq1 ivrewards2 ivsmack2 ivconseq2 ///
          obposit1 quposit1 eewarmth1, ///
ivrewards2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          ivrewards1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
          obposit2 quposit2 eewarmth2, ///
obposit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          obneg1 oblimit1 obposit2 obneg2 oblimit2 ///
          ivplay1 quposit1 eewarmth1, ///
obposit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
```

```
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          obposit1 obneg1 oblimit1 obneg2 oblimit2 ///
          ivplay2 quposit2 eewarmth2, ///
obneg1:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          obposit1 oblimit1 obposit2 obneg2 oblimit2 ///
          ivsmack1 quneg1 eecriticism1, ///
obneg2:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          obposit1 obneg1 oblimit1 obposit2 oblimit2 ///
          ivsmack2 quneg2 eecriticism2, ///
oblimit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          obposit1 obneg1 obposit2 obneg2 oblimit2 ///
          ivconseq1 qulimit1, ///
oblimit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          obposit1 obneg1 oblimit1 obposit2 obneg2 ///
          ivconseq2 qulimit2, ///
quposit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
          ivplay1 obposit1 eewarmth1, ///
quposit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          quposit1 quneg1 qulimit1 quneg2 qulimit2 ///
          ivplay2 obposit2 eewarmth2, ///
quneg1:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          quposit1 qulimit1 quposit2 quneg2 qulimit2 ///
          ivsmack1 obneg1 eecriticism1, ///
quneg2:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
```

```
          i.interven `bacthlist' `theralist' ///
          quposit1 quneg1 qulimit1 quposit2 qulimit2 ///
          ivsmack2 obneg2 eecriticism2, ///
qulimit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          quposit1 quneg1 quposit2 quneg2 qulimit2 ///
          ivconseq1 oblimit1, ///
qulimit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          quposit1 quneg1 qulimit1 quposit2 quneg2 ///
          ivconseq2 oblimit2, ///
eewarmth1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          eecriticism1 eewarmth2 eecriticism2 ///
          ivplay1 obposit1 quposit1, ///
eewarmth2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          eewarmth1 eecriticism1 eecriticism2 ///
          ivplay2 obposit2 quposit2, ///
eecriticism1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          eewarmth1 eewarmth2 eecriticism2 ///
          ivsmack1 obneg1 quneg1, ///
eecriticism2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bacthlist' `theralist' ///
          eewarmth1 eewarmth2 eecriticism1 ///
          ivsmack2 obneg2 quneg2, ///
childiq1: age i.gender loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
mumdep1:  age i.gender loneparn freemeal childiq1  ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch , ///
pacshyp2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
```

```
                pacscon1 pacscon2 ///
                i.interven `bacthlist' `theralist' , ///
      pareducn: age i.gender loneparn freemeal childiq1 mumdep1 ethminor ///
                pacscon1 pacscon2 ///
                i.rawbatch ///
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
      freemeal: age i.gender loneparn childiq1 mumdep1 ethminor ///
                pacscon1 pacscon2 ///
                i.rawbatch ///
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
      loneparn: age i.gender freemeal childiq1 mumdep1 ethminor ///
                pacscon1 pacscon2 ///
                i.rawbatch ///
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
      ethminor: age i.gender loneparn freemeal childiq1 mumdep1 ///
                pacscon1 pacscon2 ///
                i.rawbatch ///
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1) ///
    saving(linebtmi, replace) m($nimpute) seed($seed)
```

*****Step 2: Analysis the imputed data using mixed effect model and provide estimates of parameters a, b, c' and derived ab, c=ab+c', ab/ab+c'**

```
matrix coef_ind=J($nimpute,6,0)
matrix onerow=J(1,$nimpute,1)
forval m = 1/$nimpute {
set more off
use linebtmi, replace
preserve
  keep if _mj==`m'
```

**The noncons option suppresses constant term from the random-effects equation.*
**This means the random variations are all from therapy groups in the treated arm*
**i.e. therapy group variation in the control arm (interven==0 and therapy group==0) is zero*

```
    *** effect on mediator1 ***
    quietly: xtmixed `medvar'2 interven pacscon1 `medvar'1 `confvar' || rawbatch: || therapgp:interven, ///
             nocons difficult iterate(20)
    matrix M=e(b)
    local a1 = M[1,1]
    *** direct (non-mediated) intervention effect on outcome and (adjusted) effect of mediator on outcome ***
    quietly: xtmixed pacscon2 interven `medvar'2 pacscon1 `medvar'1 `confvar' || rawbatch: || therapgp:interven, ///
            nocons difficult iterate(20)
```

```
      matrix A=e(b)
      local b1 = A[1,2]
      local cprime = A[1,1]

      ***Define the mediation effect
      local prod1 = `a1'*`b1'
      local derived_c = `a1'*`b1'+`cprime'
      local med_prop1 = `a1'*`b1'/(`a1'*`b1'+`cprime')

      ***Save parameters of interest in a matrix
      matrix coef_ind[`m',1] = `a1'
      matrix coef_ind[`m',2] = `b1'
      matrix coef_ind[`m',3] = `cprime'
      matrix coef_ind[`m',4] = `prod1'
      matrix coef_ind[`m',5] = `derived_c'
      matrix coef_ind[`m',6] = `med_prop1'
restore
}
matrix coef=onerow*coef_ind/$nimpute
matrix colnames coef = a1 b1 cprime prod1 derived_c med_prop1
matlist coef
clear
  set obs 1
 *variables coefa--coefprod store the bootstrap point estimates and mediation effect
  gen repid = 0
  svmat coef, names(col)
  save coef_conf_`medvar', replace
```

*****Part 2: Generating confidence interval of the causal parameters of interest using MI-BT combined procedure**

*****Step 1: Bootstrap re-sampling**

****Give the values of computer directory path, number of imputation and seed to global variables path, number of imputations and seed respectively*

```
  gl path = "U:\PhD Parenting Trial Analysis\SPOKES "
  gl nimpute=20
  gl seed=542168375
```

*****Cluster bootstrap sampling: sampling with replacement at higher cluster level - randomization batch.*
*****This sampling strategy is recommended in section 3.8 of Davison and Hinkleys book*
*****Davison, A. C. & Hinkley, D. V. (1997) Bootstrap methods and their application,(New York, Cambridge University Press).*
*****Therapy group is at the lower level and nested within randomization batches.*
*****Note that the size of the randomization batches are not balanced, therefore different BT samples may have different number of observations.*
*****For clustered data, the sampling variance of estimates will generally depends on the number of clusters sampled.*
*****Literature: Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X. & Lai, S. (2010) Nonparametric bootstrapping for hierarchical data. Journal of Applied Statistics, 37(9), 1487-1498.*

```
set more off
set seed $seed
forvalue i=1/1000{
use "$path\SPOKES_Short_03Apr2013", clear
```

****Combine the random batches 1 and 7, 2 and 4 respectively to eliminate mini batches for operational reason.*
****The combined batches can be treated as higher level cluster compared with the mini batches*
****We sampling the combined cluster consistent with the cluster sampling at higher level strategy suggested in Davison book.*

```
replace randombatch=1 if randombatch==7
replace randombatch=2 if randombatch==4
bsample, cluster(randombatch) idcluster(newbatch)
gen newtherapgp=newbatch*1000+therapgp
save bsample`i', replace
}
```

Step 2: MICE for each BT sample

*****Multiple Imputation by Chained Equations*

```
set more off
forvalue repeat=1/1000 {
use bsample`repeat', clear
```

****Set the batch values to another variable for generating batch dummy variables for both treated and control arms*

```
gen bootsbatch=newbatch
```

****the therapy group variable in the treated arm takes account of the random batch effect and the therapy group effect.*
****set the therapy group value to 0 in the control arm*

```
replace newtherapgp=0 if interven==0
```

```
***the randomization batch in the control arm takes account of the batch effect.
***set randomization batch to 0 in the treated arm
  replace newbatch=0 if interven==1
***work out the number of therapy groups in the bootstrap sample and create dummy variables for each therapy group
  tabulate newtherapgp, gen(thergroup)
  local th=r(r)
  display `th'
***give the names of the list of the generated dummy therapy group variables to a local variable and
***exclude the control group dummy (therapgp=0), choose the first therapy group as reference group
  forvalues i=3/`th' {
  local minus=`i'-1
  local tgrp`i' `tgrp`minus'' thergroup`i'
  }
***work out the number of randomization batch groups in the bootstrap sample and create dummy variables for each randomization batch group
  tabulate newbatch, gen(batchgroup)
  local ba=r(r)
  display `ba'
***give the names of the list of the generated dummy batch group variables to a local variable and
***exclude the treated group dummy (randombatch=0), choose the first batch as the reference group
  forvalues i=3/`ba' {
  local baminus=`i'-1
  local bgrp`i' `bgrp`baminus'' batchgroup`i'
  }
***ice command - Multiple Imputation by chained equations
 ***The variables in the first row of ice command are child behaviour outcome, directly observed parenting mediators at time point 2
 ***The variables in the rows below ice are
      *1. child behaviour at baseline, directly observed parenting behaviours at baseline and
       *2. baseline characteristics to be included in the analysis model
      *3. the dummy grouping variables
      *4. the alternative measurement of child behaviour at baseline and time point 2 and the auxiliary baseline characteristics variables
      *5. the interview parenting behaviour variables at baseline and time point 2
      *6. the alternative interview parenting behaviour variables at baseline and time point 2
      *7. the questionnaire parenting behaviour variables at baseline and time point 2
      *8. the expressed emotion parenting behaviour variables at baseline and time point 2
***Use cmd option to define the regression command to be used in imputation - linear regression is used here
***Use match option to do predictive mean matching for discrete variables
***Use eq option to define customised prediction equations for each incomplete variable
***Predictor variables for child outcome of interest (i.e pacscon2):
```

**1. baseline characteristics (i.e. age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn)*
**2. baseline child outcome and alternative child outcome at baseline and time point2 (i.e. pacscon1 pacshyp1 pacshyp2)*
**3. group variables (i.e. i.interven `bgrp`ba'' `tgrp`th'')*
**4. parenting behaviour variables at time point 2 measured via interview, questionnaire, direct obseration and EE*
** (i.e. ivplay2 vsmack2 ivconseq2 quposit2 quneg2 qulimit2 obposit2 obneg2 oblimit2 eewarmth2 eecriticism2)*

****Predictor variables for parenting behaviour at baseline:*
**1. baseline characteristics (i.e. age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn for ivplay1)*
**2. child outcome at baseline and time point2 (i.e. pacscon1 pacscon2 for ivplay1)*
**3. batch variables (i.e. i.bootsbatch for ivplay1. As treatment randomization is applied to parenting trials*
** and the baseline measurements were done before receiving treatment, we assume the conditional distribution of*
** baseline variables do not depend on intervention group variable and therapy group variables given the observed values)*
**4. the other parenting behaviour variables measured by the same measurement method (i.e. ivsmack1 ivconseq1 for ivplay1) at baseline and*
** all the parenting behaviour variables measured by the same measurement method (i.e. ivplay2 ivsmack2 ivconseq2 for ivplay1) at time point 2*
**5. variables measured the same parenting behaviour but by different measurement methods at baseline*
** (alternative parenting behaviour at baseline. i.e. obposit1 quposit1 eewarmth1 for ivplay1)*
****Predictor variables for parenting behaviour mediator (time point 2)*
**1. baseline characteristics (i.e. age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn for ivplay2)*
**2. child outcome at baseline and and time point2 (i.e. pacscon1 pacscon2 for ivplay2)*
**3. group variables (i.e. i.interven `bgrp`ba'' `tgrp`th'' for ivplay2)*
**4. all the parenting behaviour variables measured by the same measurement method (i.e. ivplay1 ivsmack1 ivconseq1 for ivplay2) at baseline and*
** the other parenting behaviour variables measured by the same measurement method (i.e. ivsmack2 ivconseq2 for ivplay2) at time point 2*
**5. variables measured the same parenting behaviour but by different measurement methods at time point 2*
** (alternative parenting behaviour mediators. i.e. obposit2 quposit2 eewarmth2 for ivplay2)*
****The prediction equations for parenting behaviour variables below follow the same structures as ivplay1 and ivplay2*
****The variables below are baseline characteristics variables (i.e. childiq1 mumdep1 pareducn freemeal loneparn ethminor)*
****The predictor variables are:*
**1. the other baseline characteristics variables*
**2. child outcome at baseline and time point2 (i.e. pacscon1 pacscon2)*
**3. batch variables (i.e. i.bootsbatch. As treatment randomization is applied to parenting trials*
** and the baseline measurements were done before receiving treatment, we assume the conditional distribution of*
** baseline variables is not depend on intervention group variable and therapy group variables given the observed values)*
**4. parenting behaviour variables at baseline (i.e. ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1)*

```
ice pacscon2 obposit2 obneg2 oblimit2 ///
  pacscon1 obposit1 obneg1 oblimit1 age i.gender childiq1 loneparn freemeal mumdep1 ///
  i.interven i.bootsbatch `bgrp`ba'' `tgrp`th''  ///
```

```
pacshyp1 pacshyp2 ethminor pareducn ///
ivplay1 ivsmack1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
eewarmth1 eecriticism1 eewarmth2 eecriticism2, ///
cmd(ivplay1 ivplay2 ivsmack1 ivsmack2 ivconseq1 ivconseq2 ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
    eewarmth1 eewarmth2 eecriticism1 eecriticism2 ///
    childiq1 mumdep1 pacshyp2 quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
    obposit1 obneg1 oblimit1 obposit2 obneg2 oblimit2:regress) ///
match(ivplay1 ivplay2 ivsmack1 ivsmack2 ivconseq1 ivconseq2 ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
     eewarmth1 eewarmth2 eecriticism1 eecriticism2) ///
eq(pacscon2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
             pacscon1 pacshyp1 pacshyp2 ///
             i.interven `bgrp`ba'' `tgrp`th''  ///
             ivplay1 ivplay2 ivsmack2 ivconseq2 ///
             quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
             obposit1 obneg1 oblimit1 obposit2 obneg2 oblimit2 ///
             eewarmth1 eecriticism1 eewarmth2 eecriticism2, ///
   ivplay1:  age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
             pacscon1 pacscon2 ///
             i.bootsbatch ///
             ivsmack1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
             obposit1 quposit1 eewarmth1, ///
   ivplay2:  age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
             pacscon1 pacscon2 ///
             i.interven `bgrp`ba'' `tgrp`th''  ///
             ivplay1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
             obposit2 quposit2 eewarmth2, ///
   ivsmack1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
             pacscon1 pacscon2 ///
             ivplay1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
             obneg1 quneg1 eecriticism1, ///
   ivsmack2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
             pacscon1 pacscon2 ///
             i.interven `bgrp`ba'' `tgrp`th''  ///
             ivplay1 ivsmack1 ivconseq1 ivplay2 ivconseq2 ///
             obneg2 quneg2 eecriticism2, ///
   ivconseq1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
             pacscon1 pacscon2 ///
             ivplay1 ivsmack1 ivplay2 ivsmack2 ivconseq2 ///
             oblimit1 qulimit1, ///
```

```
ivconseq2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          ivplay1 ivsmack1 ivconseq1 ivplay2 ivsmack2 ///
          oblimit2 qulimit2 eecriticism2, ///
ivpraise1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          ivsmack1 ivconseq1 ivpraise2 ivsmack2 ivconseq2 ///
          obposit1 quposit1 eewarmth1, ///
ivpraise2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          ivpraise1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
          obposit2 quposit2 eewarmth2, ///
ivrewards1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          ivsmack1 ivconseq1 ivrewards2 ivsmack2 ivconseq2 ///
          obposit1 quposit1 eewarmth1, ///
ivrewards2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          ivrewards1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
          obposit2 quposit2 eewarmth2, ///
obposit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          obneg1 oblimit1 obposit2 obneg2 oblimit2 ///
          ivplay1 quposit1 eewarmth1, ///
obposit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          obposit1 obneg1 oblimit1 obneg2 oblimit2 ///
          ivplay2 quposit2 eewarmth2, ///
obneg1:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          obposit1 oblimit1 obposit2 obneg2 oblimit2 ///
          ivsmack1 quneg1 eecriticism1, ///
obneg2:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
```

```
          obposit1 obneg1 oblimit1 obposit2 oblimit2 ///
          ivsmack2 quneg2 eecriticism2, ///
oblimit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          obposit1 obneg1 obposit2 obneg2 oblimit2 ///
          ivconseq1 qulimit1, ///
oblimit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          obposit1 obneg1 oblimit1 obposit2 obneg2 ///
          ivconseq2 qulimit2, ///
quposit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
          ivplay1 obposit1 eewarmth1, ///
quposit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          quposit1 quneg1 qulimit1 quneg2 qulimit2 ///
          ivplay2 obposit2 eewarmth2, ///
quneg1:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          quposit1 qulimit1 quposit2 quneg2 qulimit2 ///
          ivsmack1 obneg1 eecriticism1, ///
quneg2:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          quposit1 quneg1 qulimit1 quposit2 qulimit2 ///
          ivsmack2 obneg2 eecriticism2, ///
qulimit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          quposit1 quneg1 quposit2 quneg2 qulimit2 ///
          ivconseq1 oblimit1, ///
qulimit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          quposit1 quneg1 qulimit1 quposit2 quneg2 ///
```

```
          ivconseq2 oblimit2, ///
eewarmth1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          eecriticism1 eewarmth2 eecriticism2 ///
          ivplay1 obposit1 quposit1, ///
eewarmth2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
           i.interven `bgrp`ba'' `tgrp`th''  ///
           eewarmth1 eecriticism1 eecriticism2 ///
           ivplay2 obposit2 quposit2, ///
eecriticism1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          eewarmth1 eewarmth2 eecriticism2 ///
          ivsmack1 obneg1 quneg1, ///
eecriticism2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th''  ///
          eewarmth1 eewarmth2 eecriticism1 ///
          ivsmack2 obneg2 quneg2, ///
childiq1: age i.gender loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
mumdep1:  age i.gender loneparn freemeal childiq1  ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.bootsbatch , ///
pacshyp2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven `bgrp`ba'' `tgrp`th'' , ///
pareducn: age i.gender loneparn freemeal childiq1 mumdep1 ethminor ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
freemeal: age i.gender loneparn childiq1 mumdep1 ethminor ///
          pacscon1 pacscon2 ///
          i.bootsbatch ///
          ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
loneparn: age i.gender freemeal childiq1 mumdep1 ethminor ///
          pacscon1 pacscon2 ///
```

```
                i.bootsbatch ///
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        ethminor: age i.gender loneparn freemeal childiq1 mumdep1 ///
                pacscon1 pacscon2 ///
                i.bootsbatch ///
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1) ///
    saving(linebtmi`repeat', replace) m($nimpute) seed($seed)
    }
```

Step 3: Construction of MI-ML estimators

** Create a datafile (coefstore) with reps number of observations a data file that will list the results of the bootstrap*

```
  clear
  set obs 1000
```

**variable repid records the id number of bootstrapping*

```
  gen repid = _n
```

**variables coefa--coefprod store the bootstrap point estimates and mediation effect*

```
  gen coefa1 = .
  gen coefb1 = .
  gen coefcprime = .
  gen coefprod1 = .
  gen coefderived_c = .
  gen coefmed_prop1 = .
  save coefstoreBT_conf_`medvar', replace

  foreach repeat of numlist 1/1000 {
*********************************************************
    ** Fittiing random effect mediation regression models****
*********************************************************
matrix coef_ind=J($nimpute,6,0)
matrix onerow=J(1,$nimpute,1)
forval m = 1/$nimpute {
set more off
use "$path\linebtmi`repeat'", replace
preserve
keep if _mj==`m'
```

**confvar defines the list of most important confounding variables including interaction terms to be included in analysis models*

**The noncons option suppress constant term from the random-effects equation.*

**This means the random variations are all from therapy groups in the treated arm*

**i.e. therapy group variation in the control arm (interven==0 and therapy group==0) is zero*

```
        *** effect on mediator1 ***
        quietly: xtmixed `medvar'2 interven pacscon1 `medvar'1 `confvar' || bootsbatch: || newtherapgp:interven, ///
                 nocons difficult iterate(20)
        matrix M=e(b)
        local a1 = M[1,1]

        *** direct (non-mediated) intervention effect on outcome and (adjusted) effect of mediator on outcome ***
        quietly: xtmixed pacscon2 interven `medvar'2 pacscon1 `medvar'1 `confvar' || bootsbatch: || newtherapgp:interven, ///
                 nocons difficult iterate(20)
        matrix A=e(b)
        local b1 = A[1,2]
        local cprime = A[1,1]

        ***Define the mediation effect
        local prod1 = `a1'*`b1'
        local derived_c = `a1'*`b1'+`cprime'
        local med_prop1 = `a1'*`b1'/(`a1'*`b1'+`cprime')

        ***Save parameters of interest in a matrix
        matrix coef_ind[`m',1] = `a1'
        matrix coef_ind[`m',2] = `b1'
        matrix coef_ind[`m',3] = `cprime'
        matrix coef_ind[`m',4] = `prod1'
        matrix coef_ind[`m',5] = `derived_c'
        matrix coef_ind[`m',6] = `med_prop1'
restore
}
matrix coef_avg=onerow*coef_ind/$nimpute
 ****Store bootstrap point estimate in a datafile****
  use coefstoreBT_conf_`medvar', replace
  replace coefa1 = coef_avg[1,1] if repid ==`repeat'
  replace coefb1 = coef_avg[1,2] if repid ==`repeat'
  replace coefcprime = coef_avg[1,3] if repid ==`repeat'
  replace coefprod1 = coef_avg[1,4] if repid ==`repeat'
  replace coefderived_c = coef_avg[1,5] if repid ==`repeat'
  replace coefmed_prop1 = coef_avg[1,6] if repid ==`repeat'
  save coefstoreBT_conf_`medvar', replace
  }
```

Step 4: Calculate 95% BT confidence intervals

```
***Use formula (Efron, B. (1987) Better bootstrap confidence intervals. Journal of the American Statistical Association, 82(397), 171-185.)
*** to obtain the bias corrected confidence intervals of parameters of interest.
/*define local macro end for interactive use of do file and mata command*/
local END = "end"
/*import data coefstore_eeneg from stata to mata*/
use coefstoreBT_conf_`medvar', replace
mata
      X=st_data(.,("coefa1", "coefb1", "coefcprime", "coefprod1", "coefderived_c", "coefmed_prop1"))
`END'
/*import data coef from stata to mata*/
use coef_conf_`medvar', replace
mata
      Y=st_data(.,("a1", "b1", "cprime", "prod1", "derived_c", "med_prop1"))
/*store the number of rows and columns of matrix X*/
      n=rows(X)
      m=cols(X)
/*create matrix for lower and upper CI*/
      bc_lo=J(1,m,0)
      bc_up=J(1,m,0)
/*create scalar for alphas, zcritical value*/
      alphas = 1-(1-$alpha)/2
    zcrit = invnormal(alphas)
/*create standard error vector*/
    ser=sqrt((colsum(X:^2)-(colsum(X):^2)/n)/(n-1))
/*create bias matrix*/
      vect1=J(n,1,1)
      coefbias=X:-(vect1*Y)
/*calculate bias correct confidence interval*/
      vectp=colsum(coefbias:<0)/n
      z0hat=invnormal(vectp)
    q1=z0hat+(z0hat:-zcrit)
    q2=z0hat+(z0hat:+zcrit)
    alpha1=normal(q1)
    alpha2=normal(q2)
      for (i=1; i<=m; i++) {
           coefvec=X[.,i]
           coefvec=sort(coefvec, 1)
        low=trunc(alpha1[1,i]*(n+1))
             up=trunc(alpha2[1,i]*(n+1))
```

```
            if (low<1) low=1
            if (up>n) up=n
         bc_lo[1,i]=coefvec[low,1]
            bc_up[1,i]=coefvec[up,1]
      }
/*combine the point estiamtes, standard errors and confidence intervals into one matrix*/
      ALL=Y\ser\bc_lo\bc_up
/*export the mata matrix to stata matrix*/
      st_matrix("ALL",ALL)
`END'

/*name stata matrix column names*/
matrix colnames ALL = a1 b1 cprime prod1 derived_c med_prop1
/*save stata matrix as a datasets and add variable indicating the meaning of the rows*/
clear
  set obs 4
  gen parameter = "Estimate" if _n==1
  replace parameter = "Str" if _n==2
  replace parameter = "LowCI" if _n==3
  replace parameter = "UpCI" if _n==4
svmat ALL, names(col)
save ALL_conf_`medvar', replace
```

IV-MI-BT approach mediation analysis using SPOKES trial

The IV-MI-BT approach has similar procedure to the MI-BT approach that has been illustrated in Figure 3-2 of this thesis. To demonstrate the differences between the two statistical approaches and avoid duplication, only the IV mediation model of generating the estimate of the mediation effects of interest including its imputation model is listed in this part of appendix.

```
***Give the values of computer directory path, number of imputation and seed to global variables path, number of imputations and seed respectively
  gl path = "U:\PhD Parenting Trial Analysis\SPOKES "
  gl nimpute=20
  gl seed=542168375
  gl alpha=0.9

*confvar defines the list of most important confounding variables to be included in analysis models
local confvar = "mumdep1 childiq1 loneparn gender pareducn"
*orthogonalised instrumental variables for EE criticism
local mod="trtmumdep_res trtpareducn_res thergroup3_res thergroup4_res thergroup5_res thergroup6_res ///
thergroup7_res thergroup8_res thergroup9_res thergroup10_res thergroup11_res thergroup12_res thergroup13_res thergroup14_res"
*define putative mediator to be tested
local medvar = "eecriticism"
```

*****Step 1: Apply Multiple Imputation by chained equation to the data with missing values**

```
**********************************************************************************************************
***This programme includes baseline variable and intervention interaction terms in the imputation model
***Integration effects of treatment on both EE moderators and child outcome are included in the imputation model
***Improved passive approach is applied in this situation
**********************************************************************************************************
***Give the values of computer directory path, number of imputation and seed to global variables path, nimpute and seed respectively
  gl path = "U:\PhD Parenting Trial Analysis\SPOKES\SPOKES Interaction term included MI and BT for IV analysis"
  gl nimpute=20
  gl seed=542168375
**********************************************************************************************************************************************
**************************** Note: the therapy group variables are removed from MI model of child outcome ********************************
**************************** The imputed data are only for the eecriticism IV mediation analysis ***********************************************
**********************************************************************************************************************************************

  use "$path\SPOKES_Short_03Apr2013_orthogonal", clear
      local theralist thergroup3_res thergroup4_res thergroup5_res thergroup6_res thergroup7_res thergroup8_res ///
      thergroup9_res thergroup10_res thergroup11_res thergroup12_res thergroup13_res thergroup14_res
```

***Set the batch values to another variable for generating batch dummy variables for both treated and control arms

```
gen rawbatch=randombatch
```

****There is no need of creating batch variable for control groups only, as residual coding is applied for the therapy group variable*
****In this case, orthogonalised therapy group dummy variables are independent of treatment, batch variable is available in both treated*
****and control group. We will include the dummy batch variables for both arms in the MI model and include orthogonalised therapy group*
****variables in the MI regression where is required for matching the IV assumption*

```
ice pacscon2 obposit2 obneg2 oblimit2 ///
  pacscon1 obposit1 obneg1 oblimit1 age i.gender childiq1 loneparn freemeal mumdep1 ///
  i.interven i.rawbatch `theralist'  ///
  pacshyp1 pacshyp2 ethminor pareducn ///
  ivplay1 ivsmack1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
  ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
  quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
  eewarmth1 eecriticism1 eewarmth2 eecriticism2 ///
  attendance_res trtmumdep_res trtpareducn_res, ///
  cmd(ivplay1 ivplay2 ivsmack1 ivsmack2 ivconseq1 ivconseq2 ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
      eewarmth1 eewarmth2 eecriticism1 eecriticism2 ///
      childiq1 mumdep1 pacshyp2 quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
      obposit1 obneg1 oblimit1 obposit2 obneg2 oblimit2 trtpareducn_res:regress) ///
  match(ivplay1 ivplay2 ivsmack1 ivsmack2 ivconseq1 ivconseq2 ivpraise1 ivrewards1 ivpraise2 ivrewards2 ///
        eewarmth1 eewarmth2 eecriticism1 eecriticism2 trtpareducn_res) ///
  eq(pacscon2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
               pacscon1 pacshyp1 pacshyp2 ///
               i.interven i.rawbatch ///
               ivplay1 ivplay2 ivsmack2 ivconseq2 ///
               quposit1 quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
               obposit1 obneg1 oblimit1 obposit2 obneg2 oblimit2 ///
               eewarmth1 eecriticism1 eewarmth2 eecriticism2, ///
     ivplay1:  age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
               pacscon1 pacscon2 ///
               i.rawbatch ///
               ivsmack1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
               obposit1 quposit1 eewarmth1, ///
     ivplay2:  age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
               pacscon1 pacscon2 ///
               i.interven i.rawbatch  ///
               ivplay1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
               obposit2 quposit2 eewarmth2, ///
     ivsmack1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
```

```
          pacscon1 pacscon2 ///
          i.rawbatch ///
          ivplay1 ivconseq1 ivplay2 ivsmack2 ivconseq2 ///
          obneg1 quneg1 eecriticism1, ///
ivsmack2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch  ///
          ivplay1 ivsmack1 ivconseq1 ivplay2 ivconseq2 ///
          obneg2 quneg2 eecriticism2, ///
ivconseq1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          ivplay1 ivsmack1 ivplay2 ivsmack2 ivconseq2 ///
          oblimit1 qulimit1, ///
ivconseq2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch  ///
          ivplay1 ivsmack1 ivconseq1 ivplay2 ivsmack2 ///
          oblimit2 qulimit2 eecriticism2, ///
ivpraise1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          ivsmack1 ivconseq1 ivpraise2 ivsmack2 ivconseq2 ///
          obposit1 quposit1 eewarmth1, ///
ivpraise2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          ivpraise1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
          obposit2 quposit2 eewarmth2, ///
ivrewards1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          ivsmack1 ivconseq1 ivrewards2 ivsmack2 ivconseq2 ///
          obposit1 quposit1 eewarmth1, ///
ivrewards2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          ivrewards1 ivsmack1 ivconseq1 ivsmack2 ivconseq2 ///
          obposit2 quposit2 eewarmth2, ///
obposit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
```

```
          i.rawbatch ///
          obneg1 oblimit1 obposit2 obneg2 oblimit2 ///
          ivplay1 quposit1 eewarmth1, ///
obposit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          obposit1 obneg1 oblimit1 obneg2 oblimit2 ///
          ivplay2 quposit2 eewarmth2, ///
obneg1:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          obposit1 oblimit1 obposit2 obneg2 oblimit2 ///
          ivsmack1 quneg1 eecriticism1, ///
obneg2:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          obposit1 obneg1 oblimit1 obposit2 oblimit2 ///
          ivsmack2 quneg2 eecriticism2, ///
oblimit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          obposit1 obneg1 obposit2 obneg2 oblimit2 ///
          ivconseq1 qulimit1, ///
oblimit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          obposit1 obneg1 oblimit1 obposit2 obneg2 ///
          ivconseq2 qulimit2, ///
quposit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          quneg1 qulimit1 quposit2 quneg2 qulimit2 ///
          ivplay1 obposit1 eewarmth1, ///
quposit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          quposit1 quneg1 qulimit1 quneg2 qulimit2 ///
          ivplay2 obposit2 eewarmth2, ///
quneg1:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
```

```
          quposit1 qulimit1 quposit2 quneg2 qulimit2 ///
          ivsmack1 obneg1 eecriticism1, ///
quneg2:   age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          quposit1 quneg1 qulimit1 quposit2 qulimit2 ///
          ivsmack2 obneg2 eecriticism2, ///
qulimit1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          quposit1 quneg1 quposit2 quneg2 qulimit2 ///
          ivconseq1 oblimit1, ///
qulimit2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          quposit1 quneg1 qulimit1 quposit2 quneg2 ///
          ivconseq2 oblimit2, ///
eewarmth1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          eecriticism1 eewarmth2 eecriticism2 ///
          ivplay1 obposit1 quposit1, ///
eewarmth2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch ///
          eewarmth1 eecriticism1 eecriticism2 ///
          ivplay2 obposit2 quposit2 attendance_res, ///
eecriticism1: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
          eewarmth1 eewarmth2 eecriticism2 ///
          ivsmack1 obneg1 quneg1, ///
eecriticism2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.interven i.rawbatch `theralist' ///
          eewarmth1 eewarmth2 eecriticism1 ///
          ivsmack2 obneg2 quneg2 ///
          trtmumdep_res trtpareducn_res, ///
childiq1: age i.gender loneparn freemeal mumdep1 ethminor pareducn ///
          pacscon1 pacscon2 ///
          i.rawbatch ///
```

```
                ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        mumdep1:  age i.gender loneparn freemeal childiq1  ethminor pareducn ///
                  pacscon1 pacscon2 ///
                  i.rawbatch ///
                  ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        pacshyp2: age i.gender childiq1 loneparn freemeal mumdep1 ethminor pareducn ///
                  pacscon1 pacscon2 ///
                  i.interven i.rawbatch, ///
        pareducn: age i.gender loneparn freemeal childiq1 mumdep1 ethminor ///
                  pacscon1 pacscon2 ///
                  i.rawbatch ///
                  ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        freemeal: age i.gender loneparn childiq1 mumdep1 ethminor ///
                  pacscon1 pacscon2 ///
                  i.rawbatch ///
                  ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        loneparn: age i.gender freemeal childiq1 mumdep1 ethminor ///
                  pacscon1 pacscon2 ///
                  i.rawbatch ///
                  ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        ethminor: age i.gender loneparn freemeal childiq1 mumdep1 ///
                  pacscon1 pacscon2 ///
                  i.rawbatch ///
                  ivplay1 ivsmack1 ivconseq1 obposit1 obneg1 oblimit1 quposit1 quneg1 qulimit1 eewarmth1 eecriticism1, ///
        trtmumdep_res: age i.gender loneparn freemeal childiq1 ethminor pareducn ///
                  pacscon1 pacscon2 ///
                  i.rawbatch eecriticism1 eecriticism2, ///
        trtpareducn_res: age i.gender loneparn freemeal childiq1 mumdep1 ethminor ///
                  pacscon1 pacscon2 ///
                  i.rawbatch eecriticism1 eecriticism2) ///
saving(linebtmi_ortho, replace) m($nimpute) seed($seed)
```

*****Step 2: Analysis the imputed data using mixed effect model and provide estimates of parameters a, b, c' and derived ab, c=ab+c', ab/ab+c'**

```
matrix coef_ind=J($nimpute,6,0)
matrix onerow=J(1,$nimpute,1)
forval m = 1/$nimpute {
set more off
use "$path\linebtmi_ortho", replace
preserve
  keep if _mj==`m'
```

****Use the 2SLS IV regression approach calculate b path and cprimt (the mediator effect on child outcome and the indirect treatment effect on child outcome)*

```
       *** first stage: get the predicted mediator value including IV variable in the mediator regression model***
      quietly: xtmixed `medvar'2 interven pacscon1 `medvar'1 `confvar' `mod' || rawbatch: , diff emiterate(50)
       matrix M=e(b)
       local a1 = M[1,1]
       predict `medvar'hat, fitted
       *** second stage: regress the outcome on the fitted mediator and the confounders
       quietly: xtmixed pacscon2 interven `medvar'hat pacscon1 `medvar'1 `confvar' || rawbatch: , ///
                diff emiterate(50)
matrix A=e(b)
       local b1 = A[1,2]
       local cprime = A[1,1]

       ***Define the mediation effect
       local prod1 = `a1'*`b1'
       local derived_c = `a1'*`b1'+`cprime'
       local med_prop1 = `a1'*`b1'/(`a1'*`b1'+`cprime')

       ***Save parameters of interest in a matrix
       matrix coef_ind[`m',1] = `a1'
       matrix coef_ind[`m',2] = `b1'
       matrix coef_ind[`m',3] = `cprime'
       matrix coef_ind[`m',4] = `prod1'
       matrix coef_ind[`m',5] = `derived_c'
       matrix coef_ind[`m',6] = `med_prop1'
restore
matrix coef=onerow*coef_ind/$nimpute
matrix colnames coef = a1 b1 cprime prod1 derived_c med_prop1
matlist coef
clear
  set obs 1
 *variables coefa--coefprod store the bootstrap point estimates and mediation effect
  gen repid = 0
  svmat coef, names(col)
  save coef_`medvar'_thera_edu_dep_ortho, replace
```

References

A'CAMPO, L. E., WEKKING, E. M., SPLIETHOFF-KAMMINGA, N. G., STIJNEN, T. & ROOS, R. A. 2012. Treatment effect modifiers for the patient education programme for Parkinson's disease. *International Journal of Clinical Practice,* 66**,** 77-83.

ACHENBACH, T. M. & RESCORAL, L. A. 2001. Manual for the ASEBA School-Age Forms & Profiles. *Burlington VT: University of Vermont, Research Center for Children, Youth and Families.*

AGUINIS, H. 2004. *Regression analysis for categorical moderators,* New York, Guilford Press.

AINSWORTH, M. D. S., BLEHAR, M. C., WATERS, E. & WALL, S. 1978. *Patterns of Attachment: A Psychological Study of the Strange Situation,* Hillsdale, Lawrence Erlbaum.

AITKEN, R. C. 1969. Measurement of feelings using visual analogue scales. *Proceedings of the Royal Society of Medicine,* 62**,** 989.

ALBERT, J. M. 2008. Mediation analysis via potential outcomes models. *Statistics in Medicine,* 27**,** 1282-1304.

ALEXANDERSSON, A. 2004. Graphing confidence ellipses: An update of ellip for Stata 8. *Stata Journal,* 4**,** 242-256.

AMERICAN PSYCHIATRIC ASSOCIATION 1994. *Diagnostic and Statistical Manual of Mental Disorders,* Washington, DC, American Psychiatric Association.

ANGOLD, A. & COSTELLO, E. J. 2000. The Child and Adolescent Psychiatric Assessment (CAPA). *Journal of the American Academy of Child and Adolescent Psychiatry,* 39**,** 39-48.

ANGRIST, J., IMBENS, G. & KRUEGER, A. B. 1995. Jackknife instrumental variables estimation. National Bureau of Economic Research Cambridge, Mass., USA.

ARBUCKLE, J. L. 1995–2007. *Amos™ 16.0 User's Guide,* Chicago, Amos Development Corporation.

BAGNER, D. M. & EYBERG, S. M. 2007. Parent-child interaction therapy for disruptive behavior in children with mental retardation: A Randomized controlled trial. *Journal of Clinical Child and Adolescent Psychology,* 36**,** 418-429.

BARON, R. M. & KENNY, D. A. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology,* 51**,** 1173-1182.

BAUM, C., SCHAFFER, M. & STILLMAN, S. 2003. IVREG2: stata module for extended instrumental variables/2SLS and GMM estimation (Statistical Software Component No. S425401). *Boston College*.

BAUMRIND, D. 1991. Effective parenting during the early adolescent transition. *Family transitions,* 2**,** 111-163.

BEAUCHAINE, T. P., WEBSTER-STRATTON, C. & REID, M. J. 2005. Mediators, moderators, and predictors of 1-year outcomes among children treated for early-onset conduct problems: A latent growth curve analysis. *Journal of Consulting and Clinical Psychology,* 73**,** 371-388.

BECK, A. T., WARD, C. H., MENDELSON, M., MOCK, J. & ERBAUGH, J. 1961. An inventory for measuring depression. *Archives of General Psychiatry,* 4**,** 561-571.

BECK AT, W. C. M. M. M. J. E. J. 1961. AN inventory for measuring depression. *Archives of General Psychiatry,* 4**,** 561-571.

BENTLER, P. M. & DUDGEON, P. 1996. Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology,* 47**,** 563-592.

BENTLER, P. M. & WU, E. J. C. 2005. *EQS 6.1 for Windows User Guide,* Encino, CA: Multivariate Software, Inc.

BIERMAN, K. L., COIE, J. D., DODGE, K. A., GREENBERG, M. T., LOCHMAN, J. E., MCMAHON, R. J., PINDERHUGHES, E. E. & CONDUCT PROBLEMS PREVENTION RES, G. 1999. Initial impact of the Fast Track prevention trial for

conduct problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology,* 67**,** 648-657.
BLOOMQUIST, M. L. & SCHNELL, S. V. 2002. *Helping children with aggression and conduct problems: Best practices for intervention*, Guilford Press.
BOGGS, S. R., EYBERG, S. & REYNOLDS, L. A. 1990. Concurrent validity of the Eyberg child-behavior inventory. *Journal of Clinical Child Psychology,* 19**,** 75-78.
BOUND, J., JAEGER, D. A. & BAKER, R. M. 1995. Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association,* 90**,** 443-450.
BOWLBY, J. 1969/1982. *Attachment and Loss,* New York, Basic Books.
BRETHERTON, I. & MUNHOLLAND, K. A. 1999. Internal Working Models in Attachment Relationships: A Construct Revisited. *In:* CASSIDY, J. & SHAVER, P. R. (eds.) *Handbook of Attachment: Theory, Research and Clinical Applications.* New York: Guilford Press.
BROWN, G. W., BIRLEY, J. L. & WING, J. K. 1972. Influence of family life on the course of schizophrenic disorders: a replication. *British Journal of Psychiatry,* 121**,** 241-58.
BRYANT, P. E. & BRADLEY, L. 1985. *Children's reading problems,* Oxford, UK, Basil Blackwell.
BURGESS, S. & THOMPSON, S. G. 2011. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Statistics in Medicine,* 30**,** 1312-1323.
BURGESS, S., THOMPSON, S. G. & GENETICS COLLABORATION, C. C. 2012. Methods for meta-analysis of individual participant data from Mendelian randomisation studies with binary outcomes. *Statistical Methods in Medical Research*.
BUSE, A. 1992. The bias of instrumental variable estimators. *Econometrica: Journal of the Econometric Society***,** 173-180.
CARPENTER, J. R., GOLDSTEIN, H. & KENWARD, M. G. 2011. REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types. *Journal of Statistical Software,* 45**,** 1-14.
CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. & CRAINICEANU, C. M. 2012. *Measurement error in nonlinear models: a modern perspective*, CRC press.
CASSIDY, J. & SHAVER, P. R. 2008. *Handbook of Attachment: Theory, Research and Clinical Applications,* New York, Guilford Press.
CLAY, M. 1993. *An observation survey of early literacy achievement,* Portsmouth, NH, Heinemann.
COLLINS, L. M., SCHAFER, J. L. & KAM, C.-M. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods,* 6**,** 330.
CONDUCT PROBLEMS PREVENTION RESEARCH GROUP 1999. Initial impact of the fast track prevention trial for conduct problems: I. The high-risk sample. *Journal of Consulting and Clinical Psychology,* 67**,** 631-647.
CRAGG, J. G. & DONALD, S. G. 1993. Testing identifiability and specification in instrumental variable models. *Econometric Theory,* 9**,** 222-240.
CRAIG, C. C. 1936. On the Frequency Function of xy. *The Annals of Mathematical Statistics,* 7**,** 1-15.
DAVIDSON, R. & MACHKINNON, J. G. 1993. *Estimation and inference in econometrics,* New York, Oxford University Press.
DAVISON, A. C. & HINKLEY, D. V. 1997. *Bootstrap Methods and their Application,* New York, Cambridge University Press.
DE GRAAF, I., SPEETJENS, P., SMIT, F., DE WOLFF, M. & TAVECCHIO, L. 2008a. Effectiveness of the Triple P Positive Parenting Program on behavioral problems in children - A meta-analysis. *Behavior Modification,* 32**,** 714-735.

DE GRAAF, I., SPEETJENS, P., SMIT, F., DE WOLFF, M. & TAVECCHIO, L. 2008b. Effectiveness of the Triple P Positive Parenting Program on Parenting: A Meta-Analysis. *Family Relations,* 57**,** 553-566.

DE MAESSCHALCK, R., JOUAN-RIMBAUD, D. & MASSART, D. L. 2000. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems,* 50**,** 1-18.

DENHAM, S. A., WORKMAN, E., COLE, P. M., WEISSBROD, C., KENDZIORA, K. T. & ZAHN-WAXLER, C. 2000. Prediction of externalizing behavior problems from early to middle childhood: The role of parental socialization and emotion expression. *Development and Psychopathology,* 12**,** 23-45.

DISHION, T. J., C., F. D. & PATTERSON, G. R. 1996. The development and ecology of antisocial behavior. *In:* CICCHETTI, D. & COHEN, D. J. (eds.) *Developmental Psychopathology.* John Wiley & Sons, Inc.

DONALD, S. G. & NEWEY, W. K. 2001. Choosing the number of instruments. *Econometrica,* 69**,** 1161-1191.

DOWDNEY, L., MRAZEK, D., QUINTON, D. & RUTTER, M. 1984. Observation of parent child interaction with 2-year-olds to 3-year-olds. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 25**,** 379-407.

DOWDNEY, L., SKUSE, D., RUTTER, M., QUINTON, D. & MRAZEK, D. 1985. The nature and qualities of parenting provided by women raised in institutions. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 26**,** 599-625.

DRETZKE, J., DAVENPORT, C., FREW, E., BARLOW, J., STEWART-BROWN, S., BAYLISS, S., TAYLOR, R. S., SANDERCOCK, J. & HYDE, C. 2009. The clinical effectiveness of different parenting programmes for children with conduct problems: a systematic review of randomised controlled trials. *Child and adolescent psychiatry and mental health,* 3**,** 7.

DUNN, G. & BENTALL, R. 2007. Modelling treatment-effect heterogeneity in randomized controlled trials of complex interventions (psychological treatments). *Statistics in Medicine,* 26**,** 4719-4745.

DUNN, G., EMSLEY, R., LIU, H. & LANDAU, S. 2013. Integrating biomarker information within trials to evaluate treatment mechanisms and efficacy for personalised medicine. *Clin Trials*.

DUNN, L. L. M., DUNN, L. M., WHETTON, C. & BURLEY, J. 1997. *British picture vocabulary scale second edition (BPVS-II).* Windsor, UK, NFER Nelson.

EDDY, J. M. & CHAMBERLAIN, P. 2000. Family management and deviant peer association as mediators of the impact of treatment condition on youth antisocial behavior. *Journal of Consulting and Clinical Psychology,* 68**,** 857-863.

EFRON, B. & TIBSHIRANI, R. J. 1993. *An introduction to the bootstrap*, CRC press.

ELLIOTT, C. D., SMITH, P. & MCCULLOCH, K. 1996a. *British Ability Scales,* Windsor, NFER-Nelson.

ELLIOTT, C. D., SMITH, P. & MCCULLOCH, K. 1996b. British Ability Scales second edition (BAS II): administration and scoring manual. *London: NFER-Nelson*.

EMSLEY, R., DUNN, G. & WHITE, I. R. 2010. Mediation and moderation of treatment effects in randomised controlled trials of complex interventions. *Statistical Methods in Medical Research,* 19**,** 237-270.

EMSLEY, R. & LIU, H. 2013. PARAMED: Stata module to perform causal mediation analysis using parametric regression models. *Statistical Software Components*.

EYBERG, S. M., NELSON, M. M. & BOGGS, S. R. 2008. Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child and Adolescent Psychology,* 37**,** 215-237.

EYBERG, S. M. & PINCUS, D. 1999. Child Behaviour Inventory and Sutter-Eyeberg Student Behaviour Inventory – Revised. *Psychological Assessment Resources.* Odessa.

FARRINGTON, D. P. 1995. The Twelfth Jack Tizard Memorial Lecture* The development of offending and antisocial behaviour from childhood: Key findings from the Cambridge study in delinquent development. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 36**,** 929-964.

FARRINGTON, D. P. & WELSH, B. C. 2003. Family-based prevention of offending: A meta-analysis. *Australian and New Zealand Journal of Criminology,* 36**,** 127-151.

FAY, R. E. When are inferences from multiple imputation valid. Proceedings of the Survey Research Methods Section of the American Statistical Association, 1992. 227-32.

FELLER, W. 2008. *An introduction to probability theory and its applications*, John Wiley & Sons.

FERGUSSON, D. M., JOHN HORWOOD, L. & RIDDER, E. M. 2005. Show me the child at seven: the consequences of conduct problems in childhood for psychosocial functioning in adulthood. *Journal of Child Psychology and Psychiatry,* 46**,** 837-849.

FORGATCH, M. S., PATTERSON, G. R. & DEGARMO, D. S. 2005. Evaluating fidelity: Predictive validity for a measure of competent adherence to the oregon model of parent management training. *Behavior Therapy,* 36**,** 3-13.

FOX, J. 2006. Teacher's corner: structural equation modeling with the sem package in R. *Structural Equation Modeling,* 13**,** 465-486.

FREDERICKSON, N., FRITH, U. & REASON, R. 1997. Phonological Assessment Battery (manual and test materials).

FRITZ, M. S. & MACKINNON, D. P. 2007. Required sample size to detect the mediated effect. *Psychological Science,* 18**,** 233-239.

FULLER, W. A. 1977. Some properties of a modification of the limited information estimator. *Econometrica: Journal of the Econometric Society***,** 939-953.

GARDNER, F., BURTON, J. & KLIMES, I. 2006. Randomised controlled trial of a parenting intervention in the voluntary sector for reducing child conduct problems: outcomes and mechanisms of change. *Journal of Child Psychology and Psychiatry,* 47**,** 1123-1132.

GARDNER, F., CONNELL, A., TRENTACOSTA, C. J., SHAW, D. S., DISHION, T. J. & WILSON, M. N. 2009. Moderators of Outcome in a Brief Family-Centered Intervention for Preventing Early Problem Behavior. *Journal of Consulting and Clinical Psychology,* 77**,** 543-553.

GARDNER, F., HUTCHINGS, J., BYWATER, T. & WHITAKER, C. 2010. Who benefits and how does it work? Moderators and mediators of outcome in an effectiveness trial of a parenting intervention. *Journal of Clinical Child & Adolescent Psychology,* 39**,** 568-580.

GENNETIAN, L. A., BOS, J. M. & MORRIS, P. A. 2002. Using instrumental variables analysis to learn more from social policy experiments. *New York, NY: Manpower Demonstration Research Corporation*.

GENNETIAN, L. A., MAGNUSON, K. & MORRIS, P. A. 2008. From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology,* 44**,** 381-394.

GENNETIAN, L. A., MORRIS, P. A., BOS, J. M. & BLOOM, H. S. 2005. Constructing Instrumental Variables from Experimental Data to Explore How Treatments Produce Effects. *In:* BLOOM, H. S. (ed.) *Learning more from social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

GLYMOUR, M. M., TCHETGEN, E. J. T. & ROBINS, J. M. 2012. Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions. *American Journal of Epidemiology,* 175**,** 332-339.

GOETGHEBEUR, E. & LAPP, K. 1997. The effect of treatment compliance in a placebo-controlled trial: Regression with unpaired data. *Applied Statistics-Journal of the Royal Statistical Society Series C,* 46**,** 351-364.

GOETGHEBEUR, E. & VANSTEELANDT, S. 2005. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Statistical Methods in Medical Research,* 14**,** 397-415.

GOLDBERG, D. P. 1972a. The detection of psychiatric illness by questionnaire: A technique for the identification and assessment of non-psychotic psychiatric illness.

GOLDBERG, D. P. 1972b. *The detection of psychiatric illness by questionnaire: a technique for the identification and assessment of non-psychotic psychiatric illness,* London, Oxford University Press.

GOLDBERG, D. P., GATER, R., SARTORIUS, N., USTUN, T. B., PICCINELLI, M., GUREJE, O. & RUTTER, C. 1997. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine,* 27**,** 191-197.

GOLLOB, H. F. & REICHARDT, C. S. 1991. Interpreting and estimating indirect effects assuming time lags really matter. *In:* COLLINS, L. M. & HORN, J. L. (eds.) *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* Washington, DC: American Psychological Association.

GOODMAN, R. 2001. Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry,* 40**,** 1337-1345.

GREEN, S. B. 1991. How many subjects does it take to do a regression analysis. *Multivariate Behavioral Research,* 26**,** 499-510.

GREENBERG, M. T. 1999. Attachment and psychopathology in childhood. *In:* CASSIDY, J. & SHAVER, P. R. (eds.) *Handbook of attachment.* New York: Guilford Press.

GREENBERG, M. T., SPELTZ, M. L. & DEKLYEN, M. 1993. The role of attachment in the early development of disruptive behavior problems. *Development and Psychopathology,* 5**,** 191-213.

GREENLAND, S. & MORGENSTERN, H. 2001. Confounding in health research. *Annual Review of Public Health,* 22**,** 189-212.

HAHN, J., HAUSMAN, J. & KUERSTEINER, G. 2004. Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations. *The Econometrics Journal,* 7**,** 272-306.

HAHN, J. Y. & HAUSMAN, J. 2003. Weak instruments: Diagnosis and cures in empirical econometrics. *American Economic Review,* 93**,** 118-125.

HANISCH, C., FREUND-BRAIER, I., HAUTMANN, C., JÄNEN, N., PLÜCK, J., BRIX, G., EICHELBERGER, I. & DÖPFNER, M. 2010. Detecting effects of the indicated prevention Programme for Externalizing Problem behaviour (PEP) on child symptoms, parenting, and parental quality of life in a randomized controlled trial. *Behavioural and Cognitive Psychotherapy,* 38**,** 95-112.

HANISCH, C., HAUTMANN, C., PLÜCK, J., EICHELBERGER, I. & DÖPFNER, M. 2014. The prevention program for externalizing problem behavior (PEP) improves child behavior by reducing negative parenting: analysis of mediating processes in a randomized controlled trial. *Journal of Child Psychology and Psychiatry,* 55**,** 473-484.

HANSEN, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society***,** 1029-1054.

HENRY, J. D. & CRAWFORD, J. R. 2005. The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology,* 44**,** 227-239.

HERNÁN, M. A. & ROBINS, J. M. 2015. *Causal Inference*, Chapman & Hall/CRC.

HETHERINGTON, E. M. 1999. Chapter I. Family Functioning and the Adjustment of Adolescent Siblings in Diverse Types of Families. *Monographs of the Society for Research in Child Development,* 64**,** 1-25.

HINSHAW, S. P. 2002. Intervention research, theoretical mechanisms, and causal processes related to externalizing behavior patterns. *Development and Psychopathology,* 14**,** 789-818.

HUSSONG, A. M., CURRAN, P. J. & BAUER, D. J. 2013. Integrative data analysis in clinical psychology research. *Annu Rev Clin Psychol,* 9**,** 61-89.

IBMCORP 2011. IBM SPSS Statistics for Windows, Version 20.0. Armonk, NY: IBM Corp.

IMAI, K., KEELE, L. & TINGLEY, D. 2010. A General Approach to Causal Mediation Analysis. *Psychological Methods,* 15**,** 309-334.

IMAI, K., TINGLEY, D. & YAMAMOTO, T. 2013. Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 176**,** 5-51.

IMBENS, G. W. & ROSENBAUM, P. R. 2005. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168**,** 109-126.

JO, B., STUART, E. A., MACKINNON, D. P. & VINOKUR, A. D. 2011. The use of propensity scores in mediation analysis. *Multivariate Behavioral Research,* 46**,** 425-452.

JÖRESKOG, K. G. & SÖRBOM, D. 1996 - 2001. *LISREL 8: User's Reference Guide,* USA, Scientific Software International, Inc.

JUDD, C. M. & KENNY, D. A. 1981. Process analysis - estimating mediation in treatment evaluations. *Evaluation Review,* 5**,** 602-619.

KAHAN, B. C. & MORRIS, T. P. 2013. Adjusting for multiple prognostic factors in the analysis of randomised trials. *BMC Medical Research Methodology,* 13**,** 99.

KAZDIN, A. E. 2005. *Parent management training: treatment for oppositional, aggressive, and antisocial behavior in children and adolescents,* New York, Oxford University Press

KEAR, D. J., COFFNIAN, G. A., MCKENNA, M. C. & AMBROSIO, A. L. 2000. Measuring attitude toward writing: A new tool. *The Reading Teacher,* 54**,** 1.

KLING, A., FORSTER, M., SUNDELL, K. & MELIN, L. 2010. A Randomized Controlled Effectiveness Trial of Parent Management Training With Varying Degrees of Therapist Support. *Behavior Therapy,* 41**,** 530-542.

KOBAK, R. R., COLE, H. E., FERENZGILLIES, R., FLEMING, W. S. & GAMBLE, W. 1993. Attachment and emotion regulation during mother-teen problem solving: A control theory analysis. *Child Development,* 64**,** 231-245.

KRAEMER, H. C. 2013. Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. *Statistics in Medicine,* 32**,** 1964-1973.

KRAEMER, H. C., STICE, E., KAZDIN, A., OFFORD, D. & KUPFER, D. 2001. How do risk factors work together? Mediators, moderators, and independent, overlapping, and proxy risk factors. *American Journal of Psychiatry,* 158**,** 848-856.

LA GRECA, A. M., SILVERMAN, W. K. & LOCHMAN, J. E. 2009. Moving Beyond Efficacy and Effectiveness in Child and Adolescent Intervention Research. *Journal of Consulting and Clinical Psychology,* 77**,** 373-382.

LAMBERT, P. C., SUTTON, A. J., ABRAMS, K. R. & JONES, D. R. 2002. A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology,* 55**,** 86-94.

LANCE, C. E. 1988. Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Applied Psychological Measurement,* 12**,** 163-175.

LANDAU, S. 2011. Statistical design of mental health complex intervention trials. *Trials,* 12**,** A145.

LAROSE, S. & BERNIER, A. 2001. Social support processes: mediators of attachment state of mind and adjustment in late adolescence. *Attachment and Human Development,* 3**,** 96-120.

LITTLE, R. J. 1988. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics,* 6**,** 287-296.

LITTLE, R. J. 1992. Regression with missing X's: a review. *Journal of the American Statistical Association,* 87**,** 1227-1237.

LITTLE, R. J. 1993. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association,* 88**,** 125-134.
LITTLE, R. J. & RUBIN, D. B. 2002. Statistical analysis with missing data.
LOEBER, R. & DISHION, T. 1983. Early predictors of male delinquency: a review. *Psychological Bulletin,* 94**,** 68-99.
MACKINNON, D. P. 2008. *Introduction to Statistical Mediation Analysis,* New York, Taylor & Francis Group, LLC.
MACKINNON, D. P. & DWYER, J. H. 1993. Estimating Mediated Effects in Prevention Studies. *Evaluation Review,* 17**,** 144-158.
MACKINNON, D. P. & FAIRCHILD, A. J. 2009. Current Directions in Mediation Analysis. *Current Directions in Psychological Science,* 18**,** 16-20.
MACKINNON, D. P., FAIRCHILD, A. J. & FRITZ, M. S. 2007a. Mediation analysis. *Annual Review of Psychology.*
MACKINNON, D. P. & FRITZ, M. S. 2007. Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behav Res Methods,* 39**,** 384-389.
MACKINNON, D. P., FRITZ, M. S., WILLIAMS, J. & LOCKWOOD, C. M. 2007b. Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. *Behav Res Methods,* 39**,** 384-389.
MACKINNON, D. P., LOCKWOOD, C. M., HOFFMAN, J. M., WEST, S. G. & SHEETS, V. 2002. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods,* 7**,** 83-104.
MACKINNON, D. P., LOCKWOOD, C. M. & WILLIAMS, J. 2004. Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research,* 39**,** 99-128.
MACKINNON, D. P., WARSI, G. & DWYER, J. H. 1995. A simulation study of mediated effect measures. *Multivariate Behavioral Research,* 30**,** 41-62.
MARKIE-DADDS, C. & SANDERS, M. R. 2006. Self-directed Triple P (Positive Parenting Program) for mothers with children at-risk of developing conduct problems. *Behavioural and Cognitive Psychotherapy,* 34**,** 259-275.
MARTENS, E. P., PESTMAN, W. R., DE BOER, A., BELITSER, S. V. & KLUNGEL, O. H. 2006. Instrumental variables application and limitations. *Epidemiology,* 17**,** 260-267.
MARTINEZ, C. R. & FORGATCH, M. S. 2001. Preventing problems with boys' noncompliance: Effects of a parent training intervention for divorcing mothers. *Journal of Consulting and Clinical Psychology,* 69**,** 416-428.
MCARDLE, J. J. & MCDONALD, R. P. 1984. Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology,* 37**,** 234-251.
MCCULLOCH, C. E. S. S. R. N. J. M. 2008. *Generalized, linear, and mixed models,* New Jersey, Wiley.
MCNAUGHTON, S., GLYNN, T. & ROBINSON, V. 1987. *Pause, prompt and praise: Effective tutoring for remedial reading*.
MELTZER, H., GATWARD, R., GOODMAN, R. & FORD, T. 2000. The Mental Health of Children and Adolescents in Great Britain. *London: Office for National Statistics*.
MENG, X.-L. 1994. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science***,** 538-558.
MONETTE, G. 1990. Geometry of multiple regression and interactive 3-D graphics. *Modern methods of data analysis***,** 209-256.
MUTHÉN, L. K. & MUTHÉN, B. O. 1998-2010. *Mplus User's Guide.,* Los Angeles, Muthén & Muthén.
NICE 2013. Antisocial behaviour and conduct disorders in children and young people: recognition, intervention and management. *The British Psychological Society and the Royal College of Psychiatrists,* 158**,** 208-210.
OFFICE FOR NATIONAL STATISTICS 2002. Social Focus in Brief: Ethnicity 2002.
OFFICE FOR NATIONAL STATISTICS 2005. The National Statistics Socio-economic Classification User Manual.

PALMER, T. M., LAWLOR, D. A., HARBORD, R. M., SHEEHAN, N. A., TOBIAS, J. H., TIMPSON, N. J., SMITH, G. D. & STERNE, J. A. 2012. Using multiple genetic variants as instrumental variables for modifiable risk factors. *Statistical Methods in Medical Research,* 21**,** 223-242.

PATTERSON, G. R. 1982. Coercive family process. *In:* R., P. G. (ed.) *A social learning approach.* Eugene: Castalia.

PATTERSON, G. R., DEBARYSHE, B. D. & RAMSEY, E. 1989. A developmental perspective on antisocial-behavior. *American Psychologist,* 44**,** 329-335.

PEARL, J. Direct and indirect effects. Proceedings of the seventeenth conference on uncertainty in artificial intelligence, 2001. Morgan Kaufmann Publishers Inc., 411-420.

PEVALIN, D. J. 2000. Multiple applications of the GHQ-12 in a general population sample: an investigation of long-term retest effects. *Social Psychiatry and Psychiatric Epidemiology,* 35**,** 508-512.

PINHEIRO, J. C. & BATES, D. M. 2000. *Mixed effects models in S and S-PLUS*, Springer.

PIQUERO, A. R., FARRINGTON, D. P., WELSH, B. C., TREMBLAY, R. & JENNINGS, W. G. 2009. Effects of early family/parent training programs on antisocial behavior and delinquency. *Journal of Experimental Criminology,* 5**,** 83-120.

PREACHER, K. J. & KELLEY, K. 2011. Effect Size Measures for Mediation Models: Quantitative Strategies for Communicating Indirect Effects. *Psychological Methods,* 16**,** 93-115.

REID, M. J., WEBSTER-STRATTON, C. & BAYDAR, N. 2004. Halting the development of conduct problems in head start children: The effects of parent training. *Journal of Clinical Child and Adolescent Psychology,* 33**,** 279-291.

REN, S., LAI, H., TONG, W., AMINZADEH, M., HOU, X. & LAI, S. 2010. Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics,* 37**,** 1487-1498.

REYNO, S. M. & MCGRATH, P. J. 2006. Predictors of parent training efficacy for child externalizing behavior problems - a meta-analytic review. *Journal of Child Psychology and Psychiatry,* 47**,** 99-111.

RILEY, R. D., LAMBERT, P. C. & ABO-ZAID, G. 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. *British Medical Journal,* 340**,** 521-525.

ROBINS, J. M. 2003. Semantics of causal DAG models and the identification of direct and indirect effects. *In:* GREEN, P., HJORT, N. & RICHARDSON, S. (eds.) *Highly Structured Stochastic Systems.* Oxford: Oxford University Press.

ROBINS, J. M. & GREENLAND, S. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology,* 3**,** 143-155.

ROBINS, J. M. & GREENLAND, S. 1994. Adjusting for differential rates of prophylaxis therapy for PCP in high-dose versus low-dose AZT treatment arms in an aids randomized trial. *Journal of the American Statistical Association,* 89**,** 737-749.

ROBINS, J. M. & WANG, N. 2000. Inference for imputation estimators. *Biometrika,* 87**,** 113-124.

ROBINSON, E. A. & EYBERG, S. M. 1981. The dyadic parent-child interaction coding system - standardization and validation. *Journal of Consulting and Clinical Psychology,* 49**,** 245-250.

ROBINSON, G. K. 1991. That BLUP is a good thing: The estimation of random effects. *Statistical Science***,** 15-32.

ROTHENBERG, T. J. 2007. Approximating the distributions of econometric estimators and test statistics.

RUBIN, D. B. 1976. Inference and missing data. *Biometrika,* 63**,** 581-592.

RUBIN, D. B. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association,* 75**,** 591-593.

RUBIN, D. B. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association,* 91**,** 473-489.

RUBIN, D. B. 2004. *Multiple imputation for nonresponse in surveys,* London, John Wiley & Sons.
RUTTER, M. 2005. Environmentally mediated risks for psychopathology: Research strategies and findings. *Journal of the American Academy of Child and Adolescent Psychiatry,* 44**,** 3-18.
RUTTER, M., BISHOP, D., PINE, D., SCOTT, S., STEVENSON, J. S., TAYLOR, E. A. & THAPAR, A. 2011. *Rutter's child and adolescent psychiatry,* London, John Wiley & Sons.
SANDERS, M. R. 1999. Triple P-Positive Parenting Program: towards an empirically validated multilevel parenting and family support strategy for the prevention of behavior and emotional problems in children. *Clinical Child and Family Psychology Review,* 2**,** 71-90.
SASINSTITUTEINC 2011. Base SAS® 9.3 Procedures Guide. Cary, NC: SAS Institute Inc.
SAWA, T. 1969. The Exact Sampling Distribution of Ordinary Least Squares and Two-Stage Least Squares Estimators. *Journal of the American Statistical Association,* 64**,** 923-937.
SCHAFER, J. 1997. Analysis of incomplete multivariate data. *New York: Chapman Hall*.
SCHAFER, J. L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research,* 8**,** 3-15.
SCHEFFE, H. 1999. *The analysis of variance,* London, John Wiley & Sons.
SCOTT, S. 2008. Parenting Programs. *In:* RUTTER, M., BISHOP, D., PINE, D., SCOTT, S., STEVENSON, J., TAYLOR, E. & THAPAR, A. (eds.) *Rutter's Child and Adolescent Psychiatry.* 5th ed. London: Blackwell Publishing.
SCOTT, S. & DADDS, M. R. 2009. Practitioner Review: When parent training doesn't work: theory-driven clinical strategies. *Journal of Child Psychology and Psychiatry,* 50**,** 1441-1450.
SCOTT, S., KNAPP, M., HENDERSON, J. & MAUGHAN, B. 2001a. Financial cost of social exclusion: follow up study of antisocial children into adulthood. *British Medical Journal,* 323**,** 191-194.
SCOTT, S., O'CONNOR, T. G., FUTH, A., MATIAS, C., PRICE, J. & DOOLAN, M. 2010a. Impact of a parenting program in a high-risk, multi-ethnic community: the PALS trial. *Journal of Child Psychology and Psychiatry,* 51**,** 1331-1341.
SCOTT, S., SPENDER, Q., DOOLAN, M., JACOBS, B. & ASPLAND, H. 2001b. Multicentre controlled trial of parenting groups for childhood antisocial behaviour in clinical practice. *British Medical Journal,* 323**,** 194-197.
SCOTT, S., SYLVA, K., BECKETT, C., DOOLAN, M., KALLITSOGLOU, A., BEECHAM, J. & FORD, T. 2012a. Which type of parenting programme best improves child behaviour and reading? The Helping Children Achieve trial. United Kingdom: Department of Education.
SCOTT, S., SYLVA, K., BECKETT, C., KALLITSOGLOU, A., DOOLAN, M. & FORD, T. 2012b. Should parenting programmes to improve children's life chances address child behaviour, reading skills, or both? Rationale for the Helping Children Achieve trial. *European Journal of Developmental Psychology,* 9**,** 47-60.
SCOTT, S., SYLVA, K., DOOLAN, M., PRICE, J., JACOBS, B., CROOK, C. & LANDAU, S. 2010b. Randomised controlled trial of parent groups for child antisocial behaviour targeting multiple risk factors: the SPOKES project. *Journal of Child Psychology and Psychiatry,* 51**,** 48-57.
SEAMAN, S. R. & WHITE, I. R. 2013. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research,* 22**,** 278-295.
SHEA, J. 1997. Instrument relevance in multivariate linear models: A simple measure. *Review of Economics and Statistics,* 79**,** 348-352.

SHELTON, K. K., FRICK, P. J. & WOOTTON, J. 1996. Assessment of parenting practices in families of elementary school-age children. *Journal of Clinical Child Psychology,* 25**,** 317-329.
SIMMONDS, M. C., HIGGINSA, J. P., STEWARTB, L. A., TIERNEYB, J. F., CLARKE, M. J. & THOMPSON, S. G. 2005. Meta-analysis of individual patient data from randomized trials: a review of methods used in practice. *Clinical Trials,* 2**,** 209-217.
SIMON, R. 2008. The use of genomics in clinical trial design. *Clinical Cancer Research,* 14**,** 5984-93.
SIMON, R. & ALTMAN, D. G. 1994. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer,* 69**,** 979-85.
SKRONDAL, A. & LAAKE, P. 2001. Regression among factor scores. *Psychometrika,* 66**,** 563-575.
SMALL, D. S. 2012. Mediation analysis without sequential ignorability: Using baseline covariates interacted with random assignment as instrumental variables. *Journal of Statistical Research,* 19 237-270.
SMITH, G. D. & EBRAHIM, S. 2003. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology,* 32**,** 1-22.
SMITH, G. D., EGGER, M. & PHILLIPS, A. N. 1997. Meta-analysis. Beyond the grand mean? *BMJ: British Medical Journal,* 315**,** 1610.
SOBEL, M. E. 1987. Direct and indirect effects in linear structural equation models. *Sociological Methods & Research,* 16**,** 155-176.
STAIGER, D. & STOCK, J. H. 1997. Instrumental variables regression with weak instruments. *Econometrica,* 65**,** 557-586.
STATACORP 2009. *STATAUSER'S GUIDE RELEASE 11.* Texas, A Stata Press Publication.
STATACORP 2011. *Stata Statistical Software: Release 12.* College Station: TX: StataCorp LP.
STEWART, G. B., ALTMAN, D. G., ASKIE, L. M., DULEY, L., SIMMONDS, M. C. & STEWART, L. A. 2012. Statistical analysis of individual participant data meta-analyses: a comparison of methods and recommendations for practice. *PLoS One,* 7**,** e46042.
STEWART, L. & PARMAR, M. 1993. Meta-analysis of the literature or of individual patient data: is there a difference? *The Lancet,* 341**,** 418-422.
STOCK, J. H., WRIGHT, J. H. & YOGO, M. 2002. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics,* 20**,** 518-529.
STOCK, J. H. & YOGO, M. 2002. Testing for weak instruments in linear IV regression. National Bureau of Economic Research, Inc.
STONE, C. A. & SOBEL, M. E. 1990. The robustness of estimates of total indirect effects in covariance structure models estimated by maximum. *Psychometrika,* 55**,** 337-352.
SUTTON, A. J. & HIGGINS, J. 2008. Recent developments in meta-analysis. *Statistics in Medicine,* 27**,** 625-650.
SYLVA, K., PRICE, J., CROOK, C. & ROBERTS, F. 2011. *The SPOKES literacy programme*.
SYLVA, K., SCOTT, S., TOTSIKA, V., EREKY-STEVENS, K. & CROOK, C. 2008. Training parents to help their children read: A randomized control trial. *British Journal of Educational Psychology,* 78**,** 435-455.
TAYLOR, E., SCHACHAR, R., THORLEY, G. & WIESELBERG, M. 1986. Conduct disorder and hyperactivity: I. Separation of hyperactivity and antisocial conduct in British child psychiatric-patients. *British Journal of Psychiatry,* 149**,** 760-767.
TEN HAVE, T. R., ELLIOTT, M. R., JOFFE, M., ZANUTTO, E. & DATTO, C. 2004. Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association,* 99**,** 16-25.

TEN HAVE, T. R. & JOFFE, M. M. 2010. A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research,* 21**,** 77-107.

TEN HAVE, T. R., JOFFE, M. M., LYNCH, K. G., BROWN, G. K., MAISTO, S. A. & BECK, A. T. 2007. Causal mediation analyses with rank preserving models. *Biometrics,* 63**,** 926-934.

TURNER, R. M., OMAR, R. Z., YANG, M., GOLDSTEIN, H. & THOMPSON, S. G. 2000. A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine,* 19**,** 3417-3432.

VAN BUUREN, S. & OUDSHOORN, C. 2000. *Multivariate Imputation by Chained Equations: MICE V1. 0 Users's Manual*, TNO Prevention and Health, Public Health.

VAN HEMERT, D. A., VAN DE VIJVER, F. J. R. & POORTINGA, Y. H. 2002. The Beck Depression Inventory as a Measure of Subjective Well-Being: A Cross-National Study. *Journal of Happiness Studies,* 3**,** 257-286.

VANDERWEELE, T. J. & VANSTEELANDT, S. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface,* 2**,** 457-468.

VANDERWEELE, T. J. & VANSTEELANDT, S. 2010. Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology,* 172**,** 1339-1348.

VAUGHN, C. E. 1989. Expressed emotion in family relationships - annotation. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 30**,** 13-22.

VON HIPPEL, P. T. 2009. How to impute interactions, squares, and other transformed variables. *Sociological Methodology,* 39**,** 265-291.

WAKSCHLAG, L. S., BRIGGS-GOWAN, M. J., HILL, C., DANIS, B., LEVENTHAL, B. L., KEENAN, K., EGGER, H. L., CICCHETTI, D., BURNS, J. & CARTER, A. S. 2008. Observational Assessment of Preschool Disruptive Behavior, Part II: validity of the Disruptive Behavior Diagnostic Observation Schedule (DB-DOS). *Journal of the American Academy of Child and Adolescent Psychiatry,* 47**,** 632-41.

WEBSTER-STRATTON, C. 1984. Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology,* 52**,** 666-678.

WEBSTER-STRATTON, C. 1990. Enhancing the effectiveness of self-administered videotape parent training for families with conduct-problem children. *Journal of Abnormal Child Psychology,* 18**,** 479-492.

WEBSTER-STRATTON, C. 1994. Advancing videotape parent training - a comparison study. *Journal of Consulting and Clinical Psychology,* 62**,** 583-593.

WEBSTER-STRATTON, C. & HAMMOND, M. 1997. Treating children with early-onset conduct problems: A comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology,* 65**,** 93-109.

WEBSTER-STRATTON, C. & HANCOCK, L. 1998. Training for parents of young children with conduct problems: content, methods, and therapeutic processes. *In:* BRIESMEISTER, J. & SCHAEFER, C. (eds.) *Handbook of parent training.* 2nd ed. New York: Wiley.

WEBSTER-STRATTON, C., KOLPACOFF, M. & HOLLINSWORTH, T. 1988. Self-administered videotape therapy for families with conduct-problem children: Comparison with two cost-effective treatments and a control group. *Journal of Consulting and Clinical Psychology,* 56**,** 558-566.

WEBSTER-STRATTON, C., REID, M. J. & HAMMOND, M. 2001. Preventing conduct problems, promoting social competence: A parent and teacher training partnership in Head Start. *Journal of Clinical Child Psychology,* 30**,** 283-302.

WEBSTER-STRATTON, C., REID, M. J. & HAMMOND, M. 2004. Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology,* 33**,** 105-124.

WEBSTER-STRATTON, C., REID, M. J. & STOOLMILLER, M. 2008. Preventing conduct problems and improving school readiness: evaluation of the Incredible Years Teacher and Child Training Programs in high-risk schools. *Journal of Child Psychology and Psychiatry,* 49**,** 471-488.

WECHSLER, D. 2005. *Wechsler Individual Achievement Test (WIAT-II UK)*, Harcourt Assessment.

WEERSING, V. R. & WEISZ, J. R. 2002. Mechanisms of action in youth psychotherapy. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 43**,** 3-29.

WEINBERG, C. R. 1993. Toward a clearer definition of confounding. *American Journal of Epidemiology,* 137**,** 1-8.

WEST, J., DENTON, K. & GERMINO-HAUSKEN, E. 2000. America's Kindergartners: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998-99, Fall 1998.

WHITE, I. R. & CARLIN, J. B. 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine,* 29**,** 2920-2931.

WHITE, I. R., ROYSTON, P. & WOOD, A. M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine,* 30**,** 377-399.

WOODWARD, L., DOWDNEY, L. & TAYLOR, E. 1997. Child and family factors influencing the clinical referral of children with hyperactivity: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines,* 38**,** 479-485.

WOOLDRIDGE, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data,* Cambridge, Massachusetts Institute of Technology.

WORLD HEALTH ORGANIZATION 1992. The ICD-10 Classification of Mental and Behavioural Disorders: diagnostic criteria for research. Geneva: World Health Organization.

YOUNGSTROM, E., LOEBER, R. & STOUTHAMER-LOEBER, M. 2000. Patterns and correlates of agreement between parent, teacher, and male adolescent ratings of externalizing and internalizing problems. *Journal of Consulting and Clinical Psychology,* 68**,** 1038-1050.

ZHANG, Z. Y. & WANG, L. J. 2013. Methods for Mediation Analysis with Missing Data. *Psychometrika,* 78**,** 154-184.